U.S. DEPARTMENT OF COMMERCE
National Technical Information Service
PB-284 030

A Study of System Requirements for Implementing a Chemical Substances Information Network

Sigma Data Computing Corp, Rockville, Md

Prepared for

Council on Environmental Quality, Washington, D C

Oct 77

PB 284 030

A STUDY OF SYSTEM REQUIREMENTS FOR IMPLEMENTING A CHEMICAL SUBSTANCES INFORMATION NETWORK

Draft Final Report

Prepared For

THE COUNCIL ON ENVIRONMENTAL QUALITY
ENVIRONMENTAL PROTECTION AGENCY
NATIONAL LIBRARY OF MEDICINE

Under Contract No. EQ7AC003

October 1977

SIGMA DATA COMPUTING CORP.
(301) 881-4451
12730 Twinbrook Parkway
Rockville, Maryland 20852

REPRODUCED BY
NATIONAL TECHNICAL
INFORMATION SERVICE
U. S. DEPARTMENT OF COMMERCE
SPRINGFIELD, VA. 22161

211

BIBLIOGRAPHIC DATA SHEET

1. Report No.

2.

3. PB284030

4. Title and Subtitle

A Study of System Requirements for Implementing
a Chemical Substances Information Network

5. Report Date 10-77

6.

7. Author(s)

8. Performing Organization Rept. No.

9. Performing Organization Name and Address

Sigma Data Computing Corp.
12730 Twinbrook Parkway
Rockville, Md. 20852

10. Project/Task/Work Unit No.

11. Contract/Grant No. EQ7AC003

12. Sponsoring Organization Name and Address

Council on Environmental Quality, D.C. 20006;
Environmental Protection Agency, D.C. 20460;
National Library of Medicine, Bethesda, Md. 20014

13. Type of Report & Period Covered Draft Final

14.

15. Supplementary Notes

See Mitre Corp., Chemical Substances Information Network, 2 vols. PB-274 637 and PB-274 638

16. Abstracts

The report contains a

Discussion of system requirements for the Chemical Data Bases Directory and the Chemical Structure/Nomenclature System, identification of current computer security techniques for protecting confidential data, and discussion of how the requirements of EPA's Office of Toxic Substances could be met by and integrated with the proposed Chemical Substances Information Network.

17. Key Words and Document Analysis. 17a. Descriptors

Chemical Substances Information Network
data bases
nomenclature
EPA
Toxic Substances Control Act

17b. Identifiers/Open-Ended Terms

17c. COSATI Field/Group

18. Availability Statement NTIS

19. Security Class (This Report) UNCLASSIFIED

20. Security Class (This Page) UNCLASSIFIED

21. No. of Pages 211

22. Price PC A10/MF A01

FORM NTIS-35 (REV. 10-73) ENDORSED BY ANSI AND UNESCO. THIS FORM MAY BE REPRODUCED USCOMM-DC 8265-P74

A STUDY OF SYSTEM REQUIREMENTS
FOR IMPLEMENTING A
CHEMICAL SUBSTANCES INFORMATION NETWORK

Draft Final Report

Prepared For

THE COUNCIL ON ENVIRONMENTAL QUALITY
ENVIRONMENTAL PROTECTION AGENCY
NATIONAL LIBRARY OF MEDICINE

Under Contract No. EQ7AC003

October 1977

Report No. CS-101

SIGMA DATA COMPUTING CORP.
(301) 881-4451
12730 Twinbrook Parkway
Rockville, Maryland 20852

TABLE OF CONTENTS

PAGE

1.0 INTRODUCTION 1

1.1 Objectives 2
1.2 Scope 2

2.0 CHEMICAL DATA BASES DIRECTORY 7

2.1 Introduction 7

2.1.1 Background 8
2.1.2 Approach 8

2.2 Existing Capabilities 9

2.2.1 Inventory of Federal Chemical Information Systems 9
2.2.2 Directory of Commercially Available Data Bases 12

2.3 Generalized Requirements for a Chemical Data Bases Directory 14

2.3.1 Subject Matter and Scope 14
2.3.2 Data Elements 14
2.3.3 Search and Retrieval Capabilities 22
2.3.4 User Access 24
2.3.5 User Support 24
2.3.6 Maintenance and Update 24
2.3.7 File Organization and Structure 25
2.3.8 Hardware 25
2.3.9 Software 26

2.4 Implementation of a Chemical Data Bases Directory 27

2.4.1 General System Description 27
2.4.2 Plan I - Incorporate CDBD Into MEDLARS 31
2.4.3 Plan II - Incorporate CDBD into the Existing Inventory of Federal Chemical Information Systems 36
2.4.4 Plan III - Develop CDBD Using Custom Written Software 40
2.4.5 Advantages and Disadvantages of Each Plan 43

TABLE OF CONTENTS
(Continued)

PAGE

3.0 CHEMICAL STRUCTURE/NOMENCLATURE SYSTEM 46

3.1 Introduction 46

3.1.1 Background 47
3.1.2 Approach 47

3.2 Existing Capabilities 48

3.2.1 CHEMLINE - Chemical Information On-Line 48
3.2.2 SSS/CIS - Substructure Search of the Chemical Information System 56
3.2.3 CIDS - Chemical Information and Data System 64
3.2.4 DR&D CIS - Drug Research and Development Chemical Information System 70

3.3 Generalized Requirements for a Chemical Structure/Nomenclature System 76

3.3.1 Subject Matter and Scope 76
3.3.2 Data Elements 76
3.3.3 Search and Retrieval Capabilities 80
3.3.4 User Access 83
3.3.5 User Support 83
3.3.6 Maintenance and Update 84
3.3.7 File Organization and Structure 84
3.3.8 Hardware 85
3.3.9 Software 86

3.4 A Comparison of Existing Capabilities to Generalized Requirements 87

3.5 Implementation of a Chemical Structure/Nomenclature System 90

3.5.1 General System Description 90
3.5.2 Plan I - Enhance Both CHEMLINE and SSS/CIS 92
3.5.3 Plan II - Select a Single System to Develop 98
3.5.4 Plan III - Distribute Search Capabilities Between CHEMLINE and SSS/CIS 99
3.5.5 Advantages and Disadvantages of Each Plan 102

TABLE OF CONTENTS
(Continued)

PAGE

4.0 COMPUTER SYSTEM SECURITY 104

4.1 Introduction 104

4.1.1 Background 107
4.1.2 Approach and Scope 107

4.2 Characterization of Security Risks 109

4.2.1 Introduction 109
4.2.2 Types of Security Threats 109
4.2.3 Types of Users/Intruders Posing Risks 111

4.3 Computer Technology Security Mechanisms and Techniques 113

4.3.1 Introduction 113
4.3.2 Access Control 113
4.3.3 Surveillance 119
4.3.4 Integrity Monitoring 121
4.3.5 Encryption 122
4.3.6 Software 124
4.3.7 Hardware 125

5.0 EPA/OTS REQUIREMENTS RELATED TO THE CHEMICAL SUBSTANCES INFORMATION NETWORK 132

5.1 Introduction 132

5.1.1 Background 132
5.1.2 Approach 133

5.2 TSCA Chemical Data System 135

5.2.1 General System Description 135
5.2.2 Data Base Organization and Structure 139

5.2.2.1 Proprietary Data Base 139
5.2.2.2 Public Data Base 142
5.2.2.3 Storage Estimates 146

TABLE OF CONTENTS
(Continued)

PAGE

5.2.3 Search and Retrieval Capabilities 154
5.2.4 Implementation Alternatives 157

5.2.4.1 Plan I - Dedicated Minicomputer 157
5.2.4.2 Plan II - Large Scale Time-Shared Computer 163
5.2.4.3 Security Considerations 167
5.2.4.4 Advantages and Disadvantages of Each Plan 170

5.3 OTS Interaction With CSIN 173

5.3.1 OTS Functions Related to CSIN Data Bases 174

5.3.1.1 Information 176
5.3.1.2 Problem Identification 179
5.3.1.3 Coordination and Liaison 181
5.3.1.4 Hazard Assessment 183
5.3.1.5 Health Effects 187
5.3.1.6 Ecological Effects 189
5.3.1.7 Premarket Notification 189
5.3.1.8 Regulations 193

5.3.2 Short-Term Network Description 196
5.3.3 Long-Term Network Description 198

LIST OF FIGURES

PAGE

1-1 Recommended Chemical Substances Information Network 3
2-1 Chemical Data and Information Systems Survey Form 10
2-2 Chemical Data Bases Directory System Flow Diagram 28
3-1 Chemical Structure Nomenclature System Flow Diagram 91
5-1 Existing Chemical Substances Information Network 134
5-2 TSCA Chemical Data System Information Flow Diagram 136
5-3 Proprietary TSCA CDS Data Base Structure 140
5-4 Public TSCA CDS Data Base Structure 144
5-5 Public TSCS CDS Extract Flow Chart 145(a)
5-6 TSCA CDS Indexed Data Elements 155
5-7 Functional Areas Within OTS 175
5-8 OTS Information Group 177
5-9 OTS Problem Identification Group 180
5-10 OTS Coordination and Liaison Group 182
5-11 OTS Hazard Assessment Group 184
5-12 OTS Health Effects Group 188
5-13 OTS Ecological Effects Group 190
5-14 OTS Premarket Notification Group 191
5-15 OTS Regulations Group 194
5-16 Short-Term Data Base Network 197
5-17 Long-Term Data Base Network 199

LIST OF TABLES

PAGE

2-1 Data Elements Comprising a Chemical Data Bases Directory 15
2-2 CDBD Plan I Implementation Tasks 32
2-3 CDBD Plan II Implementation Tasks 37
2-4 CDBD Plan III Implementation Tasks 41
2-5 A Comparison of CDBD Implementation Plans 44
3-1 CHEMLINE Data Elements 49
3-2 SSS/CIS Data Elements 57
3-3 CIDS Data Elements 65
3-4 DR&D CIS Data Elements 71
3-5 Data Elements Comprising a Chemical Structure/Nomenclature System 77
3-6 A Comparison of Existing Chemical Structure and Nomenclature Systems to Generalized Requirements 86
3-7 CSNS Plan I Implementation Tasks 94
3-8 CSNS Plan III Implementation Tasks 100
3-9 A Comparison of CSNS Implementation Plans 103
5-1 TSCA CDS Data Element Size 147
5-2 TSCA Chemical Data System Storage Estimates 152
5-3 TSCA CDS Plan I Implementation Tasks 159
5-4 TSCA CDS Plan II Implementation Tasks 165
5-5 A Comparison of TSCA CDS Implementation Plans 171

LIST OF WORKS CONSULTED 129

1.0 INTRODUCTION

The Toxic Substances Control Act (TSCA), which became effective January 1, 1977, provides the Environmental Protection Agency (EPA) with broad powers to regulate chemicals in commerce not covered by existing Federal regulatory authorities. The purpose of this Act is to prevent unreasonable risks of injury to health or the environment associated with the manufacture, processing, distribution in commerce, use, or disposal of chemical substances. Under TSCA, manufacturers, processors, exporters, and importers will be required to report on information concerning new chemical substances, production and use data for selected existing substances, and health and safety data required by EPA. Confidential information reported under TSCA will be protected against unauthorized disclosure.

The Act directs EPA to establish an efficient system to collect, store, retrieve, and disseminate toxicological and other scientific data that will be useful in carrying out decision making activities of TSCA. The information collected by EPA will be available to other Federal, state, and local agencies as well as the private sector to enable them to meet their requirements under TSCA.

Section 25(b) of TSCA requires the Council on Environmental Quality (CEQ) to coordinate a study of the feasibility of establishing:

- A standard classification system for chemical substances and related substances
- A standard means for storing and for obtaining rapid access to information on these systems

In order to fulfill this section of the Act, the Mitre Corporation was contracted to conduct a feasibility study to determine user requirements and systems development option's for establishing a Chemical Substances Information Network (CSIN). The Mitre study discussed the following topics:

- Identification and characterization of groups that have regulatory responsibilities to control toxic substances
- Identification and evaluation of potential information sources to satisfy user requirements
- Matching user requirements with existing systems and clarifying those areas where requirements are not being met and identifying new files to be established

- System development options for an integrated CSIN
- Administrative and physical procedures for handling confidential data

This study is an extension of the Mitre report and is undertaken through a joint effort of CEQ, EPA, and the National Library of Medicine (NLM).

1.1 Objectives

This study will analyze the system development recommendations presented in the Mitre report and determine system requirements for implementing the proposed CSIN. System requirements will include:

- file or data base specifications
- file or data base integration
- software characteristics
- hardware characteristics
- maintenance and update
- security
- processing characteristics
- manpower
- costs
- short and long term implementation alternatives
- network integration

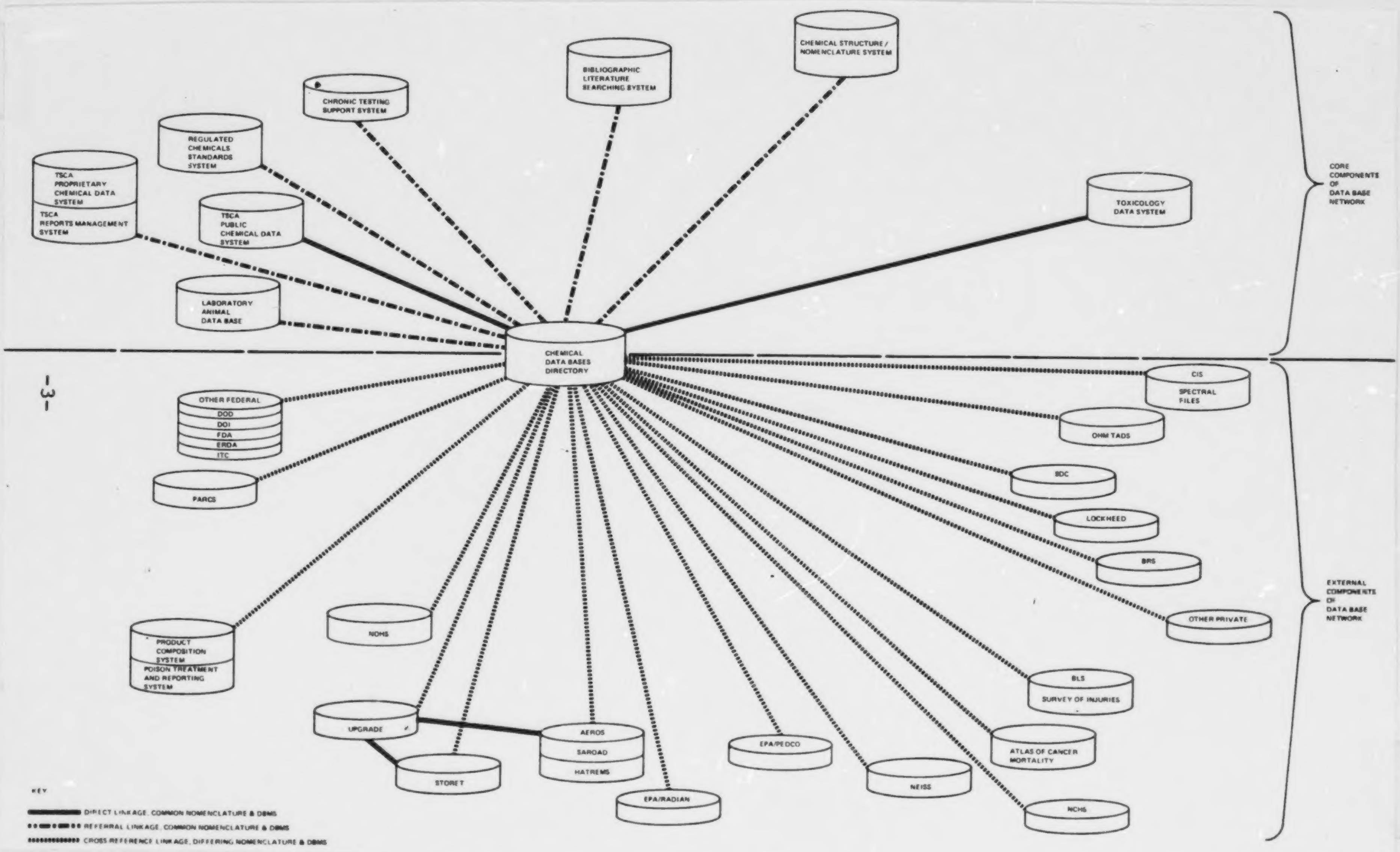
System requirements for each of the three development options presented in the Mitre report will be discussed and alternative time-phased system implementation plans presented in order to provide flexibility due to changing EPA policies and priorities. The recommendations in this report will reflect state-of-the-art technology but will also take into account budgetary, manpower, and time constraints. This study is intended to be used as a guide for agencies in making intelligent computer system implementation decisions with regard to TSCA.

1.2 Scope

The CSIN proposed in the Mitre report would serve all agencies involved in the study and regulation of chemical substances. The Network consists of existing systems and systems that must be developed or integrated. Figure 1-1 is taken from the Mitre report and illustrates the proposed CSIN. Ten systems are designated as critical "core components" and would provide the majority of the information required by OTS to carry out its responsibilities under TSCA. The ten core components include:

FIGURE 1-1

RECOMMENDED CHEMICAL SUBSTANCES INFORMATION NETWORK


CORE COMPONENTS OF DATA BASE NETWORK
EXTERNAL COMPONENTS OF DATA BASE NETWORK
CHEMICAL DATA BASES DIRECTORY
TSCA PROPRIETARY CHEMICAL DATA SYSTEM
TSCA REPORTS MANAGEMENT SYSTEM
REGULATED CHEMICALS STANDARDS SYSTEM
CHRONIC TESTING SUPPORT SYSTEM
BIBLIOGRAPHIC LITERATURE SEARCHING SYSTEM
CHEMICAL STRUCTURE / NOMENCLATURE SYSTEM
TOXICOLOGY DATA SYSTEM
TSCA PUBLIC CHEMICAL DATA SYSTEM
LABORATORY ANIMAL DATA BASE
OTHER FEDERAL
DOD
DOI
FDA
ERDA
ITC
PARCS
PRODUCT COMPOSITION SYSTEM
POISON TREATMENT AND REPORTING SYSTEM
NOHS
UPGRADE
STORET
AEROS
SAROAD
HATREMS
EPA/RADIAN
EPA/PEDCO
NEISS
ATLAS OF CANCER MORTALITY
NCHS
BLS
SURVEY OF INJURIES
OTHER PRIVATE
BRS
LOCKHEED
SDC
OHM TADS
CIS
SPECTRAL FILES
KEY
DIRECT LINKAGE, COMMON NOMENCLATURE & DBMS
REFERRAL LINKAGE, COMMON NOMENCLATURE & DBMS
CROSS REFERENCE LINKAGE, DIFFERING NOMENCLATURE & DBMS

- Chemical Data Base Directory (CDBD)
- Chemical Structure/Nomenclature System (CSNS)
- TSCA Chemical Data System (Proprietary)
- TSCA Chemical Data System (Public)
- TSCA Reports Management System
- Toxicology Data System
- Chronic Testing Support System
- Bibliographic Literature Scanning System
- Laboratory Animal Data System
- Regulated Chemicals Standards System

The first two systems can be developed independently and do not depend on EPA reporting requirements. The remainder of the systems depend on the information gathering policies of EPA. The Mitre report presents three scenarios based on various information gathering policies and plans of EPA. System development options are discussed under each scenario. The first scenario assumes EPA will collect site specific production information for all chemicals on the TSCA Inventory (50,000 - 100,000 chemicals). It also assumes that information on use, exposure, and biological effects will be collected for a selected number of chemicals (1,000 - 2,000) under Sections 8(b) and 8(a)(2) of TSCA. Scenario II assumes that EPA will collect information on use, users, and exposure for all chemicals in the TSCA Inventory and that data collected under 8(a)(2) will be increased to include 5,000 to 10,000 chemicals. The third scenario assumes that EPA will exercise its full reporting requirements under 8(a)(2) to include all chemicals in the TSCA Inventory. It is also assumed that significant new use reporting under Section 5(a) will be implemented.

This report discusses system requirements for the Chemical Data Bases Directory and the Chemical Structure/Nomenclature System, identifies current computer security techniques for protecting confidential data, and discusses how the requirements of the Office of Toxic Substances could be met by and integrated with the proposed Chemical Substances Information Network.

Section 1.0 of this study, discusses system requirements for implementing the most critical component of CSIN, the Chemical Data Base Directory, Based on the recommendations in the Mitre report, Sigma Data developed a comprehensive set of system requirements for CDBD that include the following categories:

- Subject Matter and Scope
- Data Elements
- Search and Retrieval Capabilities

- User Access
- User Support
- Maintenance and Update
- File Organization and Structure
- Hardware
- Software

Two existing systems were analyzed to determine the degree to which CDBD requirements could be met; the Inventory of Federal Chemical Information Systems, and the Directory of Commercially Available Data Bases. Based on the analysis, three alternative plans were developed to implement the Chemical Data Bases Directory. Each plan involved using the Inventory of Federal Chemical Information Systems as a basis for the Directory. The plans include incorporating the CDBD into the MEDLARS system at NLM, incorporating the CDBD into the existing Inventory of Federal Chemical Information Systems, and developing the CDBD using custom written software. Each plan includes a discussion of the tasks involved in implementation, time frames for implementation, costs and manpower requirements. The advantages and disadvantages of each plan are summarized to aid in actual implementation decisions.

Section 3.0 discusses system requirements for implementing the second critical component of CSIN, the Chemical Structure/Nomenclature System. Based on the recommendations of the Mitre report, Sigma Data developed a comprehensive set of system requirements for CSNS based on the same set of categories as CDBD. Four existing systems were analyzed:

- CHEMLINE - Chemical Information On-Line
- SSS/CIS - Substructure Search of the Chemical Information System
- CIDS - Chemical Information and Data System
- DR&D CIS - Drug Research and Development Chemical Information System

A subjective quantitative comparison of these systems to the generalized system requirements was developed to determine their applicability to CSNS. Two systems, CHEMLINE and SSS/CIS were found to satisfy most of the requirements of the CSNS. Three alternative plans were developed to implement the Chemical Structure/Nomenclature System using one or both of these systems. The plans include enhancing both CHEMLINE and SSS/CIS, selecting a single system to develop, and distributing search capabilities between CHEMLINE and SSS/CIS. Each plan includes a discussion of the tasks involved in implementation, time frames for implementation, costs, and manpower requirements. The advantages and disadvantages of each plan are summarized to aid in actual implementation decisions.

Section 4.0 identifies hardware and software related security problems associated with protecting confidential information in the proposed on-line, interactive system for a CSIN, and identifies measures that could be taken to prevent or overcome these security problems. In Section 5.0, these security mechanisms are referenced when discussing security procedures for the proprietary TSCA Chemical Data System.

Section 5.0 discusses the needs of the Office of Toxic Substances as they relate to the Chemical Substances Information Network. The discussion concentrates on the core components of the CSIN illustrated in Figure 1-1. The system development recommendations for the core components discussed in the Mitre report indicates that full implementation of the CSIN will be an evolutionary process that could span as many as ten years. Meaningful user requirements for some of the core components are as yet undefined and require further feasibility studies within the next few years to determine actual implementation requirements. These systems which include the Toxicology Data System, Bibliographic Literature Searching System, the Chronic Testing Support System, the Laboratory Animal Data System, and the Regulated Chemical Standards System, were not considered within the scope of this report.

Two other core components, the Chemical Data Bases Directory and the Chemical Structure/Nomenclature system are discussed in Sections 2.0 and 3.0. Section 5.0 discusses the remaining core components which include the proprietary and public TSCA Chemical Data System and the TSCA Reports Management System. The design of the latter system was incorporated into the TSCA CDS.

The system requirements for the TSCA Chemical Data System were developed and include a general system description, data base organization and structure, storage estimates, and search and retrieval capabilities. Two alternative implementation plans were developed which include developing the TSCA CDS on a dedicated minicomputer or a large time-shared computer. The tasks required to implement each alternative were developed along with time frames, manpower and costs. Security considerations and advantages and disadvantages were discussed to aid in implementation decisions.

The functional groups within OTS were identified and their activities as they relate to other functional groups and the component systems of the Chemical Substances Information Network were discussed. Finally, the short-term and long-term requirements of a data base system network were discussed. The network would be used by the Office of Toxic Substances to obtain information related to TSCA from the core components of the Chemical Substances Information Network.

2.0 CHEMICAL DATA BASES DIRECTORY

2.1 Introduction

This section discusses system requirements and implementation plans for developing a system that will provide on-line access to a Chemical Data Bases Directory (CDBD). Such a system will contain detailed descriptive information about the content of both computerized (machine-readable) and manual (books, microfilm, etc.) data bases. Data bases to be included in the Directory are those that contain information about the study or regulation of chemical substances.

The CDBD will be queried in order to identify those data bases that meet the user's information needs. Since the Directory will be available on-line, a user would receive immediate responses to queries. Query sessions will be conducted using a terminal and a question-and-answer dialogue between the user and the computer program performing the search and retrieval. Examples of typical questions or commands that could be posed by a user include:

- What data bases can be searched using molecular formula fragments?
- What data bases can be accessed on-line?
- What data bases contain information about environmental effects?
- What data bases contain information useful to biochemists?
- Display a complete description of CHEMLINE.
- What data bases are available through the Lockheed network?
- What data elements are searchable in TOXLINE?
- What data bases contain epidemiology data and can be searched using CAS registry number?

2.1.1 Background

The need for a Chemical Data Bases Directory was identified in a previous study completed by the Mitre Corporation of user requirements and system development options for a Chemical Substances Information Network (CSIN). The CDBD will play a key role in the overall CSIN proposed in the study. It will act as a focal point for users desiring information about other data bases in the information network. Information that could be obtained includes data base identification, location, data elements, number of records, searching capability, access procedures, subject matter, and coverage. Once the desired information is obtained, the CDBD will directly link or refer the user to the desired data bases.

Development of a CDBD has been given highest priority in the study, and it has been recommended that work begin as soon as possible. The user requirements for a CDBD are identical for each development option discussed in the study, and as such are not dependent on the various TSCA reporting requirements associated with the Chemical Substances Information Network. Therefore, implementation of CDBD can be accomplished independently of other components in CSIN.

2.1.2 Approach

In order to develop system requirements for a Chemical Data Bases Directory, a number of existing capabilities were identified by Sigma Data and an analysis of their features was conducted to determine what information content and processing procedures would be desirable in such an on-line system.

Based on the recommendations of the Mitre report, a comprehensive set of system requirements was defined for a Chemical Data Bases Directory. The requirements were divided into nine categories: subject matter and scope, data elements, search and retrieval capabilities, user access, user support, maintenance and update, file organization and structure, hardware, and software.

To insure that the best solution to implement the CDBD is adopted, three alternative implementation plans were developed after analyzing the system requirements and existing capabilities. These plans will be used as guidelines in selecting the actual implementation strategy for the Chemical Data Bases Directory.

All three plans included utilizing the existing Inventory of Federal Chemical Informations Systems as a basis for establishing the Chemical Data Bases Directory. Each plan includes and discusses the tasks involved in implementation and considers manpower, costs, and time frames associated with development of the Directory. A comparison of the various plans is presented through a discussion of their advantages and disadvantages.

2.2 Existing Capabilities

This section describes two systems that are representative of a directory of data bases. The first is a machine-readable data base that identifies chemical information systems within the Federal government; the second is a publication which documents commercially available data bases, both privately and federally owned. Each of these systems is described below, together with their relative success in meeting the requirements of the proposed Chemical Data Bases Directory.

2.2.1 Inventory of Federal Chemical Information Systems

Section 25(b) of the Toxic Substances Control Act (TSCA) mandated that the Council on Environmental Quality (CEQ) conduct a study to determine the feasibility of establishing a standard means for storing and for obtaining rapid access to information about toxic substances.

As a result, CEQ conducted a survey of data bases only within the Federal government. Both computerized (on-line or batch) and manual (books, surveys, etc.) data bases were included in the survey. To be considered in the inventory, a data base had to contain information about chemical substances.

A two-page form was sent out to various government agencies requesting a description of their chemical information systems. Figure 2-1 shows an example of the survey form. The information content of the survey form covers seven major areas of descriptive information:

- Identification
- Subject Matter and Scope
- Data Elements
- User Access
- Update
- Hardware
- Software

The information on the completed forms was converted into a computerized file using Infodata Systems' Inquire, a data base management system (DBMS) at the National Library of Medicine (NLM). Each numbered or lettered item of information on the survey form was converted to a data element in the inventory data base. Each record in the inventory includes the data elements associated with a particular base. As evidenced

SURVEY OF CHEMICAL DATA & INFORMATION SYSTEMS

IMPORTANT: PLEASE READ COVER SHEET
PRINT WRITTEN RESPONSES
CHECK ALL APPLICABLE BOXES

MAIL TO: COUNCIL ON ENVIRONMENTAL QUALITY
722 JACKSON PLACE, N.W.
WASHINGTON, D. C. 20006

1. SYSTEM IDENTIFICATION:

A. SYSTEM NAME ____
B. RESPONSIBLE OFFICE ____
AGENCY ____
DEPARTMENT ____
INFORMATION TELEPHONE (/ / /) / / - / / / / x / /
C. FORMER NAMES ____
D. MAILING ADDRESS FOR USER REQUESTS:
OFFICE ____ ROOM NO. ____
AGENCY ____
STREET ADDRESS ____
CITY ____ STATE ____ ZIP ____

2. SERVICES AVAILABLE TO USER:

- ☐ ANSWER ENQUIRIES
- ☐ REFERRAL
- ☐ LITERATURE SEARCHES
- ☐ ABSTRACTING
- ☐ DUPLICATION
- ☐ ON-SITE REFERENCING
- ☐ TAILORED SEARCHES
- ☐ DATA ANALYSIS
  - ☐ STATISTICAL ANALYSIS
  - ☐ GRAPHICAL DISPLAY
- ☐ OTHER (SPECIFY) ____

3. PURPOSE OF SYSTEM: (% OF OVERALL PURPOSE)

- ☐ MANAGEMENT & PLANNING ____
- ☐ RESEARCH & DEVELOPMENT ____
- ☐ SURVEILLANCE & MONITORING ____
- ☐ PRODUCT REGISTRATION ____
- ☐ LEGAL, LEGISLATIVE OR REGULATORY ____
- ☐ OTHER (SPECIFY) ____ ____

4. SYSTEM USERS: (% OF ALL USERS)

- ☐ PUBLIC ____
- ☐ IN-HOUSE ____
- ☐ OTHER GOVERNMENT AGENCIES ____
- ☐ OTHER (SPECIFY) ____

5. TYPE OF DATA IN SYSTEM (% OF SYSTEM DATA BASE)

- ☐ RAW ____
- ☐ ANALYZED ____
- ☐ SUMMARIZED ____
- ☐ BIBLIOGRAPHIC ____
- ☐ REFERRAL ____
- ☐ OTHER (SPECIFY) ____ ____

6. SYSTEM SOURCES: (% OF SYSTEM SOURCED)

- ☐ BOOKS ____
- ☐ JOURNALS ____
- ☐ REPORTS ____
- ☐ MONITORING ____
- ☐ MANUFACTURER ____
- ☐ SURVEYS ____
- ☐ OTHER SYSTEMS ____
- ☐ OTHER (SPECIFY) ____ ____

7. SUBJECT LIST (CHECK ANY SUBJECT ON WHICH YOUR SYSTEM COLLECTS INFORMATION OR DATA)
ALSO, PLEASE ATTACH LIST OF DATA FIELDS UTILIZED IN EACH DATA BASE
* IF NOT DISCLOSABLE, LIST IN COLUMN (1) THE % CONFIDENTIAL, IN COLUMN (2) THE % TRADE SECRET

A. CHEMICAL IDENTIFICATION/RETRIEVAL SCHEME: * (1) (2)

- ☐ PHYSICAL PROPERTIES __ __
- ☐ MOLECULAR FORMULA __ __
- ☐ CHEMICAL ABSTRACTS REGISTRY NUMBER __ __
- ☐ IUPAC __ __
- ☐ OTHER CHEMICAL NAME(S) __ __
- ☐ COMMON OR TRADE NAME(S) __ __
- ☐ WISWESSER LINE NOTATION __ __
- ☐ CHEMICAL STRUCTURE (SPECIFY) __ __
- ☐ OTHER (SPECIFY) ____ __ __

B. PRODUCT CHARACTERIZATION: * (1) (2)

- ☐ COMPOSITION __ __
- ☐ FORMULATION __ __
- ☐ MOLECULAR FORMULA __ __
- ☐ BYPRODUCTS, IMPURITIES __ __
- ☐ PHYSICAL PROPERTIES __ __
- ☐ RADIOACTIVITY __ __
- ☐ CHEMICAL PROPERTIES __ __
- ☐ DEGRADATION PRODUCTS __ __
- ☐ ENVIRONMENTAL TRANSPORT & FATE __ __
- ☐ OTHER (SPECIFY) ____ __ __

C. COMMERCIAL ASPECTS: * (1) (2)

- ☐ PRODUCTION __ __
- ☐ PLANT LOCATION __ __
- ☐ USERS __ __
- ☐ USERS (WITH AMOUNTS) __ __
- ☐ USES __ __
- ☐ USES (WITH AMOUNTS) __ __
- ☐ PLACE OF USE __ __
- ☐ WORK FORCE __ __
- ☐ SUBSTITUTES __ __
- ☐ CONTROL OPTIONS __ __
- ☐ PRODUCTION PROCESS __ __
- ☐ ECONOMIC INFORMATION __ __
- ☐ OTHER (SPECIFY) ____ __ __

D. EFFECTS INFORMATION: * (1) (2)

- ☐ BIOACCUMULATION __ __
- ☐ BIOCHEMICAL STUDIES __ __
- ☐ CLINICAL STUDIES __ __
- ☐ TOXICOLOGICAL STUDIES __ __
  - ☐ RODENTS ☐ ACUTE
  - ☐ WILDLIFE ☐ CARCINOGENICITY
  - ☐ PRIMATES ☐ MUTAGENICITY
  - ☐ IN VITRO ☐ TERATOGENICITY
  - ☐ OTHER ☐ OTHER CHRONIC
- ☐ EPIDEMIOLOGY – GENERAL __ __
- ☐ EPIDEMIOLOGY – OCCUPATIONAL __ __
- ☐ ECOLOGICAL EFFECTS __ __
- ☐ WEATHER MODIFICATION __ __
- ☐ MATERIALS __ __
- ☐ OTHER (SPECIFY) ____

E. MONITORING: * (1) (2)

- AIR ☐ QUALITY __ __
  - ☐ EMISSIONS __ __
  - ☐ WORKPLACE __ __
  - ☐ OTHER INDOOR __ __
- WATER ☐ SURFACE __ __
  - ☐ GROUND __ __
  - ☐ MARINE OR ESTUARINE __ __
  - ☐ DRINKING WATER __ __
- SOIL ☐ __ __
- PLANTS ☐ __ __
- WILDLIFE ☐ __ __
- PRODUCTS ☐ FOOD __ __
  - ☐ DRUGS __ __
  - ☐ COSMETICS __ __
  - ☐ PESTICIDES __ __
  - ☐ CONSUMER PRODUCTS __ __
  - ☐ RADIOACTIVE MATERIALS __ __
  - ☐ OTHER (SPECIFY) __ __

F. LEGAL/REGULATORY: * (1) (2)

- JURISDICTION/ACTOR ☐ INDUSTRY/COMPANY __ __
  - ☐ FEDERAL __ __
  - ☐ STATE __ __
  - ☐ LOCAL __ __
- LITIGATION ☐ __ __
- UTILIZATION REQUIREMENTS ☐ LABELING __ __
  - ☐ PERMITTING/LICENSING __ __
  - ☐ USES __ __
  - ☐ PROCESS __ __
  - ☐ STORAGE __ __
  - ☐ TRANSPORTATION __ __
  - ☐ SPILLS __ __
- COMPLIANCE ☐ __ __
- ENFORCEMENT ACTIONS ☐ __ __
- OTHER (SPECIFY) ☐ ____ __ __

A. DOES YOUR SYSTEM INTERACT WITH OTHER SYSTEMS IN SOME WAY, i.e., EXCHANGE OF DATA, AND/OR COORDINATION OF DATA COLLECTION? ☐ YES ☐ NO

B. IF "YES," IS RELATIONSHIP ☐ FORMAL OR ☐ INFORMAL? GIVE NAME OF THE OVERALL SYSTEM:

C. LIST PARTICIPATING SYSTEM MEMBERS:

(SYSTEM NAME)	(AGENCY)	(BUREAU)	(MODE OF EXCHANGE)

PROVIDE A PARAGRAPH DESCRIPTIVE OF SYSTEM (THE MATERIAL FURNISHED HERE WILL BE PUBLISHED AS THE INTRODUCTION TO YOUR SYSTEM IN OUR DIRECTORY. CONSEQUENTLY, IT IS IMPORTANT THAT THIS QUESTION BE ANSWERED IN AN INFORMATIONAL MANNER; i.e., SYSTEM PURPOSE, LEGAL BASIS (INCLUDING LEGAL CITATIONS), AND PRIMARY USERS. IF THERE IS A PUBLISHED SYSTEM ABSTRACT AVAILABLE, PLEASE ATTACH AND REFERENCE HERE).

. SYSTEM STATISTICS:

EARS ON LINE ____

EARS OFF LINE ____

SPONSE TIME PER REQUEST ____

EQUESTS PER YEAR ____

ERAGE COST PER REQUEST ____

STEM BUDGET FY 1975 ____

STEM BUDGET FY 1976 ____

TIMATED BUDGET FY 1977 ____

. OF FULL TIME EMPLOYEES ____

. OF PART TIME EMPLOYEES ____

OGRAPHIC COVERAGE: (CHECK AS MANY AS APPLY)

- ☐ LOCAL/STATE
- ☐ REGIONAL
- ☐ NATIONAL
- ☐ INTERNATIONAL

11. LANGUAGES USED:

- ☐ FORTRAN ____
- ☐ COBOL ____
- ☐ RPG ____
- ☐ BASIC ____
- ☐ PL/1 ____
- ☐ ASSEMBLY ____
- ☐ MIS OR DBMS (SPECIFY) ____
- ☐ OTHER (SPECIFY) ____

12. STORAGE MEDIA:

- ☐ PUNCH CARD ____
- ☐ MAGNETIC TAPE ____
- ☐ DISC ____
- ☐ DRUM ____
- ☐ MICROFICHE ____
- ☐ REPORT ____
- ☐ OTHER (SPECIFY) ____ ____

13. OUTPUT MEDIA:

- ☐ PUBLICATIONS ____
- ☐ COMPUTER PRINTOUT ____
- ☐ MAGNETIC TAPE ____
- ☐ MICROFICHE ____
- ☐ MICROFILM ____
- ☐ PUNCHED CARD ____
- ☐ COMPUTER GENERATED GRAPHICS ____
- ☐ OTHER (SPECIFY) ____ ____

14. COMPUTERS:

	(1)	(2)	(3)
MANUFACTURER			
MODEL			
CORE SIZE			

15. MEANS OF ACCESS TO SYSTEM:

- ☐ DIAL UP TO COMPUTER
- ☐ COMPUTER TO COMPUTER
- ☐ ON-LINE INTERACTIVE TERMINALS
- ☐ REMOTE JOB ENTRY
- ☐ MAILED INPUT
- ☐ TELEPHONE REQUEST
- ☐ WRITTEN REQUEST
- ☐ MAILING LIST
- ☐ OTHER (SPECIFY) ____

16. FILE UPDATE FREQUENCY:

- ☐ ON-LINE
- ☐ DAILY
- ☐ WEEKLY
- ☐ MONTHLY
- ☐ QUARTERLY
- ☐ ANNUALLY
- ☐ OTHER (SPECIFY) ____

. INPUT METHOD:

- ☐ KEY PUNCH ____
- ☐ KEY TO DISC ____
- ☐ KEY TO TAPE ____
- ☐ OTHER (SPECIFY ____ ____

18. STORAGE SIZE OF TOTAL DATA BASE:

FOR MERGED DATA BASES, PLEASE LIST PARENT DATA BASES :

NAME	#OF RECORDS	AS OF DATE

ANTICIPATED GROWTH (# OF CHEMICALS/YR) ____

NUMBER OF UNIQUE CHEMICALS IN FILE ____

. INFORMATION ON PERSON COMPLETING THIS FORM (THIS QUESTION MUST BE COMPLETED):

NAME: ____ POSITION: ____

AGENCY: ____

PHONE NO. (_ _ _) _ _ _ _ _ _ _ X _ _ _

OTE: PLEASE ATTACH ANY LITERATURE RELATED TO YOUR SYSTEM TO THIS SURVEY FORM. ALSO, LIST PUBLICATIONS ROUTINELY AVAILABLE TO PUBLIC OR GOVERNMENT AGENCIES.

by the form, some data elements have been assigned multiple values and certain data elements were transformed into inverted lists to provide rapid retrieval. However, Inquire allows the user to search against any data element, even if it is not inverted. Searching non-indexed data elements is more time-consuming; however, since the inventory only identifies 176 data bases, the extra search time is relatively short.

Access to the inventory of data bases is through an on-line terminal using the search and reporting capabilities of Inquire. The user can search against any data element in the inventory data base using an English-like query language. Any number of data elements can be combined in a single search using logical connectives (and, or, not, etc.) to create complex query structures. Once a query has been satisfied, any or all of the data elements describing a data base can be displayed on the terminal. A disadvantage with this DBMS is that the search and retrieval language is not geared to the inexperienced user, i.e., there is no "user-prompting". A three day class is required to learn the search and reporting capabilities.

A published version of the inventory will soon be available which will contain extensive indexing and cross-referencing to aid the user in manual searching.

2.2.2 Directory of Commercially Available Data Bases

A directory of data bases is available in a manual form through the publication Computer-Readable Bibliographic Data Bases; A Directory and Data Sourcebook. Machine-readable data bases included in this publication must be:

- composed of bibliographic information or bibliographic-related information
- commercially available or publicly available
- designed or used primarily for information retrieval purposes rather than for library processing services

The publication contains information about 301 data bases covering a broad range of subject categories. Included are such categories as aerospace, agriculture, business, chemistry, engineering, electronics, energy, environment, life sciences, medicine, physics, and social sciences. About 100 of the data bases in the publication deal with chemistry and chemical substances. Only about 60% of the data bases are produced in the United States; the remainder are produced abroad.

The publication is divided into six major parts:

- A description of each data base
- Major subject category index (50 categories)
- Name/acronym/synonym index of data base names
- Producer index
- Processor index
- Data element codes used to describe information about the data base

A user can identify a particular data base in four different ways:

- major subject category
- name/acronym/synonym of a data base
- name of the producer
- name of the processor

The description of each data base includes detailed information in ten major categories:

- Basic Information
- Producer/Distributor/Generator
- Availability and Charges
- Subject Matter and Scope
- Subject Analysis/Indexing
- Bibliographic Data Base Elements
- Tape Specifications
- Search Programs
- Data Base Services Offered
- User Aids Offered by Producer

Data for this publication are entered on-line into a data base mapping file. Data entries are then printed out, verified, or corrected on-line. Pertinent data are selected from the file and entered into a file called "Data Base of Data Bases". The data are then reformatted, printed out, corrected, and edited on-line until they are error free. Camera-ready copy is printed on a ComData terminal and submitted for offset printing. This publication is updated semiannually, and the subscription cost is $68.00 a year.

2.3 Generalized Requirements for a Chemical Data Bases Directory

This section discusses the information content, data base structure, processing, hardware, and software requirements for implementing a comprehensive Chemical Data Bases Directory. The specific requirements were determined by analyzing the conceptual requirements for CDBD presented in the Mitre report. They are not intended to be all-inclusive, but rather as a guideline to be used to determine detailed specifications for the actual Directory implemented in the CSIN.

2.3.1 Subject Matter and Scope

The on-line Directory will contain descriptive information about the maximum allowable data bases. Data bases included in the CDBD may be either manual or computerized. A data base is defined as any source from which pertinent data can be retrieved in a systematic fashion, including books, surveys and journals, as well as files on microfiche, microfilm, and computerized files on tape or disk.

Commercially available and privately owned data bases provided by the Federal government, institutions, and private industry will be included in the Directory, together with data bases containing limited access confidential information and data bases containing publicly accessable information. The coverage will be restricted to data bases containing information about the regulation and/or study of chemical substances. The CDBD will contain sufficient information to:

- aid the user in identifying data bases containing the desired type of information;
- describe search and retrieval capabilities; and,
- describe access procedures.

2.3.2 Data Elements

The data elements associated with the CDBD will provide the user with detailed information about the identification, information content, and accessability of each data base. Table 2-1 identifies the necessary data elements and identifies those that can be searched by a user. Each data element is defined below.

ACRONYM - Indicates the commonly accepted acronym or short name currently used to identify the data base, i.e., TOXLINE or EMIC.

TABLE 2-1

DATA ELEMENTS COMPRISING
A
CHEMICAL DATA BASES DIRECTORY

ELEMENT NAME	SEARCHABLE
1. BASIC IDENTIFICATION	
ACRONYM	YES
FULL NAME	YES
FORMER ACRONYM	YES
FORMER FULL NAME	YES
PRODUCER'S NAME	YES
PRODUCER'S ADDRESS	NO
GENERATOR'S NAME	YES
GENERATOR'S ADDRESS	NO
DISTRIBUTOR'S NAME	YES
DISTRIBUTOR'S ADDRESS	NO
COMPUTERIZED/MANUAL	YES
ON-LINE/BATCH SEARCH	YES
2. RELATIONSHIP TO OTHER DATA BASES	
PROVIDE INFORMATION	NO
OBTAIN INFORMATION	NO
DIRECT LINK	NO
3. INFORMATION CONTENT OF DATA BASE	
SUBJECT MATTER AND SCOPE	NO
SUBJECT CATEGORY	YES
USER COMMUNITY	YES
DATA ELEMENTS:	
SEARCHABLE AND PRINTABLE	YES
ONLY SEARCHABLE	YES
ONLY PRINTABLE	YES
DATA ELEMENT COVERAGE	NO
TIME SPAN	NO
UPDATE FREQUENCY	NO
SOURCES	NO
SIZE	NO
GROWTH	NO

TABLE 2-1
(Continued)

ELEMENT NAME	SEARCHABLE
4. ACCESS TO THE DATA BASE	
MEANS OF ACCESS	YES
PROCEDURES	NO
AVAILABILITY	NO
USE CHARGES	NO
ACCOUNT INFORMATION	NO
COST PER REQUEST	NO
YEARLY USAGE	NO
MAXIMUM NUMBER OF USERS	NO
RESPONSE TIME	NO
TERMINALS SUPPORTED	YES
TERMINAL CONSTRAINTS	NO
OUTPUT MEDIA	YES
CONFIDENTIALITY RESTRICTIONS	NO
NETWORK ACCESS	YES
USER SUPPORT	NO
USER AIDS	NO
USER SERVICES	NO
5. COMPUTER RELATED INFORMATION	
STORAGE MEDIA	NO
SEARCH AND RETRIEVAL SOFTWARE	NO
COMPUTERS USED	NO
AVAILABLE COPIES	NO
6. MISCELLANEOUS	
ONGOING MODIFICATIONS	NO
PLANNED MODIFICATIONS	NO

FULL NAME - Indicates the commonly accepted complete or full name currently used to identify the data base, i.e., Chemical Information On-Line for CHEMLINE.

FORMER ACRONYM - Lists all of the acronyms by which the data base was formerly identified, i.e., TOXICON was a former acronym for TOXLINE.

FORMER FULL NAME - Lists all of the full names by which the data base was formerly identified, i.e., Toxicology Information Conversation On-Line Network was a former name for Toxicology Information On-Line.

PRODUCER'S NAME - The name of the organization with financial responsibility and general administrative responsibility for deciding what will be included or covered by the data base. For the Federal or state government this should include the office, agency, and department.

PRODUCER'S ADDRESS - Includes a telephone number the user can call to obtain information.

GENERATOR'S NAME - The organization with responsibility for physically creating the data base. For the Federal or state government, this should include the office, agency, and department.

GENERATOR'S ADDRESS - Includes a telephone number the user can call to obtain information.

DISTRIBUTOR'S NAME - The organization with responsibility for making the data base available to the user. For the Federal or state government, this should include the office, agency, and department.

COMPUTERIZED/MANUAL - Indicates whether the data base exists in a computerized form (machine-readable disk or tape) or only in a manual form (publication, cards, etc.). If it exists in both forms, the computerized indication will be used.

ON-LINE/BATCH SEARCH - For a computerized data base, this field indicates whether a search and retrieval can be conducted using an on-line interactive mode or only through a batch type mode.

PROVIDE INFORMATION - Identifies all data bases that are processed to provide information for this data base. For instance, TOXLINE is processed to obtain all the chemical names and/or CAS registry numbers which are then used to create records in CHEMLINE. The particular elements or text describing what information is provided by the other data bases should be included in this field.

OBTAIN INFORMATION - Identifies all data bases that use this data base to obtain information. For instance, TOXLINE is used by CHEMLINE to obtain chemical names and/or CAS registry numbers. The particular elements or text describing what information is obtained in this data base by the other data bases should also be included in this field.

DIRECT LINK - Lists all the other data bases that participate in a direct link with this data base. Direct link refers to on-line systems where the user has the capability of automatically switching from one data base to another data base without the need to dial-up the other data base.

SUBJECT MATTER AND SCOPE - A brief abstract or narrative description of the data base. The description will include system purpose, reason for existence, subject matter, uses, content, type of coverage, type of data, etc.

SUBJECT CATEGORY - One of the most critical fields in the CDBD. Since many users will access the Directory to determine what data bases contain information about a particular subject, a standard set of descriptors or vocabulary will be developed to identify each subject category. A predefined hierarchical system of headings and subheadings, similar to the Medical Subject Headings (MESH) developed at the National Library of Medicine (NLM), will provide the user with a uniform subject category search capability. Before a data base is entered into the Directory, an Information Specialist will analyze the data base to determine the appropriate subject categories to be entered into this field.

USER CATEGORY - The end user groups most likely to benefit from a search of the data base. The particular groups are predefined and indexed to provide a search capability. As with Subject Category, an Information Specialist will analyze each data base to determine the appropriate user groups to be included. Typical user groups might include toxicologists, health professionals, biologists, etc.

DATA ELEMENTS - Another critical field in the Directory. Many users will access CDBD to determine what data elements are searchable and/or printable. Since data element naming conventions differ widely between data bases, a standard set of data element names will be developed to provide a uniform search capability. Differences between conventions used in various data bases will be resolved into a standard format by an Information Specialist before the data base is entered into the Directory. Data elements such as linear notation and line formula will be reduced to a standard identification, such as component line formula. A data element like chemical name will be specified in more detail to indicate the type of name provided. Chemical name descriptions will include trade names, synonyms, IUPAC, 9th Collective Index, 8th Collective Index, etc.

Elements not explicitly in a data base record, but which are searchable will be included. For instance, the user may be able to search for molecular fragments; the fragments are not actually in a data base record, but were derived from the molecular formula when the data base record was created and placed on an index file for search purposes.

Data elements are divided into three categories: searchable and printable; only searchable; only printable. The user desiring information about a data element which is only printable will be able to identify the data elements which can be used to search the data base in order to obtain the printable information.

Whenever a data base element name does not correspond to the standard name, it will be cross-referenced to the standard name and printed whenever the standard name is printed. This will indicate to the user what data element name to use when referencing the data base.

DATA ELEMENT COVERAGE - All the data elements that do not occur in all of the data base records, and the percentage of coverage and the reason for the limitation. For instance, if structural connection tables were only obtained for substances added after a specific date, the field would indicate the element name, percentage of records containing the connection tables, and an appropriate explanation for this situation.

TIME SPAN - Indicates appropriate cut off dates for including information in the data base and a brief explanation for this situation.

UPDATE FREQUENCY - The frequency with which the data base is updated (annually, monthly, quarterly, as needed, etc.)

SOURCES - The sources are used to obtain information for the data base, including books, journals, reports, monitoring data, surveys, other data bases, etc.

SIZE - The number of chemicals, citations, etc., that are contained in the data base.

GROWTH - The anticipated growth in the number of chemicals, citations, etc. over the next year.

MEANS OF ACCESS - The methods of retrieving information from the data base, including on-line interactive terminals, remote job entry, telephone request, written request, etc. Access categories are a predefined and controlled set of codes, and indexed to provide a search capability.

PROCEDURES - A detailed narrative description of the procedures required for a user to access the data base. For instance, if the user is required to telephone in a request, the narrative would describe what types of information the user would have to supply, what types of answers would be supplied, and the telephone number to call.

AVAILABILITY - The time of day and particular days of the week the user may have access to the data base.

USE CHARGES - The cost for accessing the data base. Charges include subscription fees, terminal connect fees, or charges per request.

ACCOUNT INFORMATION - If a user is required to have an account in order to access the data base, this field describes the procedures. These procedures include cost, forms to be completed, appropriate address or telephone number, contact person, etc.

COST PER REQUEST - The average cost a user might expect to pay for an answer to a request.

YEARLY USAGE - The number of requests for information (searches) made against the data base in the last year.

MAXIMUM NUMBER OF USERS - For on-line interactive data bases, this field indicates how many users can access the data base simultaneously. This should be a practical limit and not a theoretical maximum limit due to response time.

RESPONSE TIME - The average time required for a user request to be satisfied. For on-line interactive data bases, this field also indicates any unusual terminal use patterns that might exist to increase user response time.

TERMINALS SUPPORTED - Identifies the terminals that can be used to access the data base. Generic and/or specific terminal types may be indicated, i.e., TTY compatible, graphics, TEKTRONIX, etc.

TERMINAL CONSTRAINTS - Indicates any non-standard or special terminal features required to receive information from the data base. For instance, if the data base displays chemical structures using special chemical symbols, the receiving terminal would have to be equipped with special chemical type keys or fonts.

OUTPUT MEDIA - Lists the various ways a user may receive an answer to a request. These include computer printout, computer graphics, magnetic tape, microfilm, microfiche, cards, etc. This field contains predefined, controlled entries that are indexed to provide a search capability.

CONFIDENTIALITY RESTRICTIONS - Indicates what part of the data base contains confidential information which may be restricted to certain users. It also includes conditions and procedures for qualified users to follow in order to have information released.

NETWORK ACCESS - Lists each network that provides access to the data base. These include Lockheed, SDC, Tymeshare, etc. The entries are predefined, controlled, and indexed to provide a search capability.

USER SUPPORT - The name and telephone number of the person to call when a user has trouble accessing the data base or structuring a search and retrieval.

USER AIDS - Identifies user manuals, data base documentation, training sessions, etc., available to the user to aid in accessing the data base.

USER SERVICES - Indicates services related to data base searching available to the user. These may include literature searches, abstracting, tailored searches, data analysis, etc.

STORAGE MEDIA - Identifies the media on which a copy of the data base exists. Media include disk, tape, cards, microfilm, microfiche, publication, etc.

SEARCH AND RETRIEVAL SOFTWARE - Description of computer programs used to search and retrieve information. The description includes the name of the software package, developer or owner of the software, programming languages, main frame used, core required, peripheral devices, and contact person that can provide other information.

COMPUTERS USED - Describes all the computers that the software is currently operating on. The description includes manufacturer, model, operating system, and amount of core.

AVAILABLE COPIES - Indicates if a user can copy, purchase, or rent copies of the data base. It also provides the required information for obtaining the data base.

ONGOING MODIFICATIONS - Indicates any enhancements or changes that are currently being done and specifies the time of completion.

PLANNED MODIFICATIONS - Indicates any enhancements or changes that are planned and the estimated time of completion.

2.3.3 Search and Retrieval Capabilities

Search and retrieval of CDBD information will be accomplished through a question-and-answer dialogue between a user entering information on a terminal and a computer program accessing the Directory.

The following Directory data elements will be searchable by the user:

- Acronym
- Full Name
- Former Acronym

- Former Full Name
- Producer's Name
- Generator's Name
- Distributor's Name
- Computerized/Manual
- On-Line/Batch Search
- Subject Category
- User Category
- Data Elements
- Means of Access
- Terminals Supported
- Output Media
- Network Access

The user will be able to combine these data elements in a single query structure using logical connectives (AND, OR, NOT) to form a complex query. CDBD will provide the user the ability to search for words or groups of words "contained" in the full name, former full name, producer's name, generator's name, and distributor's name. With such a capability, a user will be able to display all data bases that contain the word "Chemical" in the data base name.

The results of each query will be "saved" in a temporary file. A number of "save" files will be available for each query session. Each "save" file will be uniquely identified so that they can be referenced or displayed any time during the search and retrieval session. The user will be able to combine "save" files with one another or with other data elements using logical connectives to form another query.

The user will be able to display the descriptive information about each data base. Since the description of a data base can be quite long, the user will be able to specify what data elements will be displayed as the result of a query.

An inexperienced user will be able to take advantage of appropriate "help" and "explanation" commands that will describe processing and information problems that might arise during a query session. These commands will display a description of the Directory and its data elements, a list of the standardized subject and user categories, a list of standardized data element names, and a description of each of the available commands used for search and retrieval. The search and retrieval system will provide a significant amount of user prompting to guide and aid the user through the query session.

2.3.4 User Access

The CDBD will be available to as many users as possible. This includes Federal and state government, universities, institutions, and private industry. Since the user requires immediate response, access will be through interactive terminals. To serve the anticipated number of users, the Directory will be available through a network system, either Federally or privately owned. Ideally, the CDBD should be on the same network as the other core components of the Chemical Substances Information Network to facilitate direct linkage with other data bases more easily.

2.3.5 User Support

The CDBD support group will provide the user with a comprehensive manual that describes the Directory data elements and the search and retrieval capabilities. It will include the list of standardized subject and user categories as well as the standardized data elements names. It will show representative sets of examples that illustrate the range of the search and retrieval capabilities available to the user. Since the system will be designed to respond to the inexperienced user, formal training sessions will not be necessary, although short demonstrations will be available to the user.

2.3.6 Maintenance and Update

The CDBD will not be a static system; new data bases will be added and existing data bases will be enhanced. Since these type of changes do not occur often, a semi-annual update of the Directory will be sufficient. To accommodate this update, a turn-around document consisting of a computer printout of the existing information about a data base will be sent to the producer. The producer will make any necessary information changes on the same form and return it to the CDBD group. An Information Specialist will analyze the returned form and code any required changes for input to the computer program to update the Directory. The Information Specialist will also be responsible for any changes to the standardized lists of subject categories, user categories, and data element names.

As new data bases are identified for inclusion in the Directory, an Information Specialist will be responsible for coding the required input to the computer program to add the data base to the Directory.

Since updates will be infrequent and minimal, a batch-type update will be sufficient.

2.3.7 File Organization and Structure

The CDBD will be organized such that each record contains the descriptive information for a particular data base. Each record will contain the data elements described in Section 2.3.2. Some data elements, such as Subject Category and Means of Access, will be fields containing more than one value, i.e., repeated value fields. A structural capability will be provided to search the Directory for those data elements specified as searchable in Section 2.3.3. This will necessitate an indexed, table entry, or inverted type file structure. Since the number of records and the amount of data in each record is small, a file organization could be employed that requires a complete regeneration of the data base for each update.

In case the Directory will have to be transferred to another computer, the organization should be generalized and not restricted to peculiarities of a particular machine. Use of a commercially available data base management system would be acceptable in order to provide flexibility of change and ease of implementation.

2.3.8 Hardware

Hardware requirements for the CDBD are minimal due to the data base size and processing procedures. Most large scale computers and associated operating systems will adequately fulfill the needs of CDBD. A minicomputer will be considered if it meets certain requirements, including:

- higher-level language support
- core size
- access speed
- data base management system support

Some minicomputers do not support full versions of higher level languages, such as COBOL and Fortran, which will be considered for developing the necessary software. Computer programs to process the on-line conversational search and retrieval capabilities discussed in Section 2.3.3 typically, require a minimum of 200,000 bytes of storage. These storage requirements may exhaust the limits of some minicomputers. The access speed of minicomputers is also slower than that of large scale computers, which will affect response time when trying to serve many simultaneous users. A DBMS will be considered in developing CDBD, although the number of such systems available on a minicomputer with the necessary supportive features is minimal.

Any hardware selected must provide rapid telecommunication functions to process a large number of simultaneous users. Users accessing CDBD will not require a sophisticated terminal, so any TTY-compatible terminal will be supported.

2.3.9 Software

Any software, such as update, search and retrieval, developed for the Directory should be written in a high-level language. Such languages as COBOL, Fortran, or PL/1 will be acceptable. These higher-level languages provide easier maintenance and some degree of transferability if the Directory were to be implemented on a number of different computers. It would be desirable for search and retrieval software to be re-entrant in order to process the large number of user requests simultaneously. All software will be written in modular form, which will facilitate adding new functions or changing existing functions. Input-output functions will be separated from other functions in order to isolate any machine dependent code.

To increase access speed, input-output modules may be written in assembly language. Although this language is machine dependent, it takes advantage of particular hardware characteristics and provides a faster response time for search and retrieval.

If a DBMS is used, consideration will be given to using an English-like query facility that may be provided with the system. It must be kept in mind that most of these DBMS query facilities require a great deal of training and expertise on the part of the user and do not provide much of the specific prompting required by the Directory search and retrieval system.

2.4 Implementation of a Chemical Data Bases Directory

Based on the general requirements discussed in Section 2.3 and existing capabilities discussed in Section 2.2, Sigma Data developed three alternatives for implementing a Chemical Data Bases Directory. Since some of the information required for implementing the CBDB has already been collected through the Inventory of Federal Chemical Information Systems, it was decided to use this data base to initially build the CDBD in each of the three alternatives. The three implementation plans include:

- Incorporating the CDBD into MEDLARS
- Incorporating the CDBD into the existing Inventory of Federal Chemical Information Systems
- Developing the CDBD using custom written software

The following subsections include a generalized system information flow description of the CDBD, a detailed discussion of each implementation plan, and a summary of the advantages and disadvantages of each plan.

2.4.1 General System Description

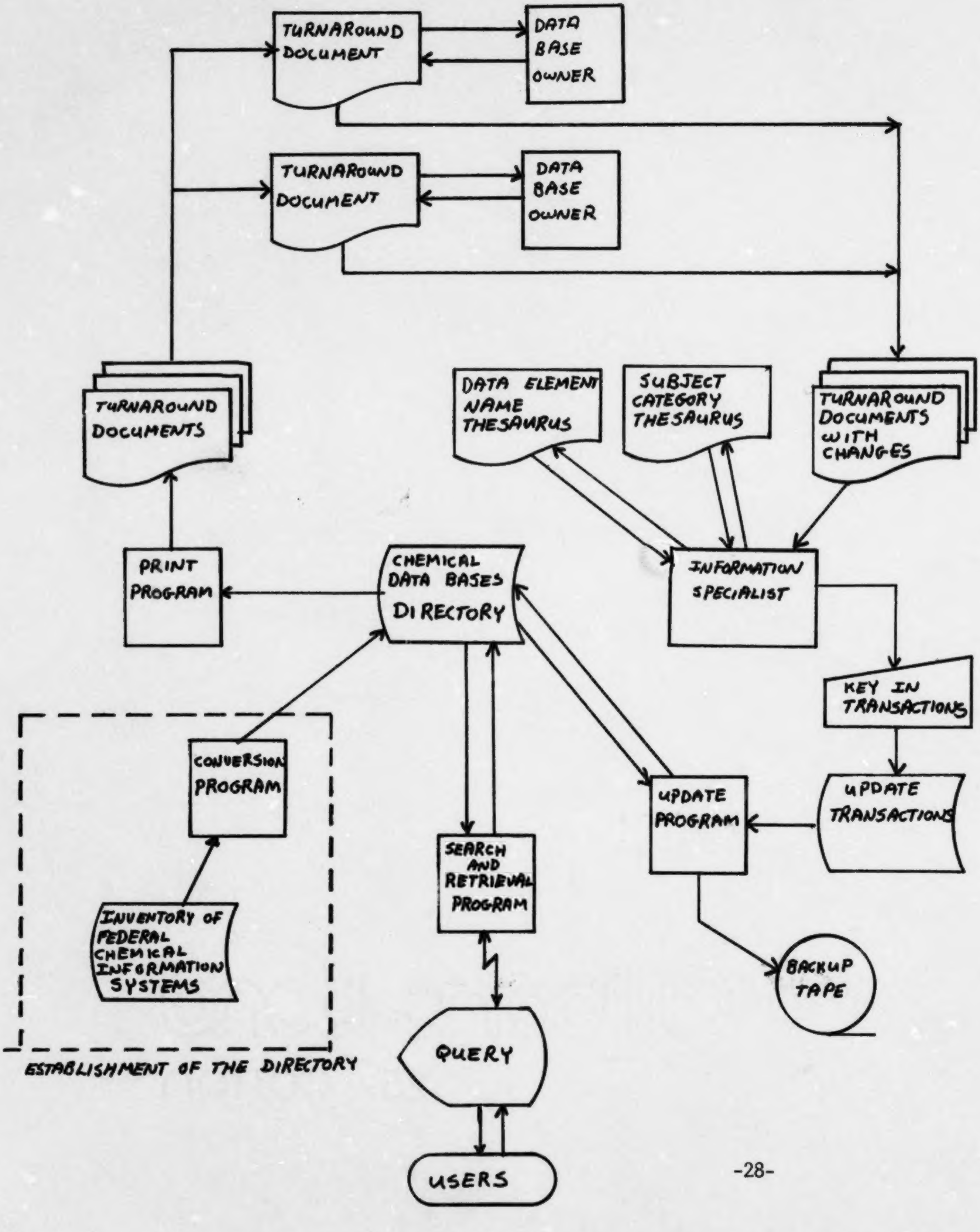
Figure 2-2 illustrates a generalized system flow diagram of the Chemical Data Bases Directory that is not dependent on the implementation plan adopted. The CDBD would initially be established by converting appropriate information existing in the Inventory of Federal Chemical Information Systems into the format and structure developed for the Directory.

Since the existing Inventory does not contain all the information required by the CDBD, more data would have to be obtained from the agencies responsible for the data bases. This would be accomplished by designing a turnaround document that identifies each data element defined for the Directory and provides space to print the contents of each data element as it is found in the Directory. The turnaround document would also contain appropriate space to add, change, or delete information about each data element. The purpose of this type of a turnaround document would be to minimize the work required to update the information about a data base.

A computer program would be developed to print the description of each data base on the turnaround documents using the information of the CDBD. The initial printing of these turnaround documents

FIGURE 2-2

CHEMICAL DATA BASES DIRECTORY SYSTEM FLOW DIAGRAM


TURNAROUND DOCUMENT
DATA BASE OWNER
TURNAROUND DOCUMENT
DATA BASE OWNER
TURNAROUND DOCUMENTS
DATA ELEMENT NAME THESAURUS
SUBJECT CATEGORY THESAURUS
TURNAROUND DOCUMENTS WITH CHANGES
PRINT PROGRAM
CHEMICAL DATA BASES DIRECTORY
INFORMATION SPECIALIST
KEY IN TRANSACTIONS
CONVERSION PROGRAM
UPDATE PROGRAM
UPDATE TRANSACTIONS
SEARCH AND RETRIEVAL PROGRAM
INVENTORY OF FEDERAL CHEMICAL INFORMATION SYSTEMS
BACKUP TAPE
QUERY
ESTABLISHMENT OF THE DIRECTORY
USERS

would contain a description of the data elements that were converted from the existing Inventory of Federal Chemical Information Systems, but would not contain a description of the new data elements developed for CDBD. To obtain the required data about the new data elements, information specialists would conduct personal interviews with responsible agencies, using the turnaround document as a guideline. This procedure would enable the information specialist to explain the purpose and desired information for each data element to the various data base representatives, and establish a good working relationship for future updates. When new data based are identified for inclusion in the CDBD, a personal interview would be conducted using a blank turnaround document.

Since the organization and structure of most data bases remains relatively stable, an annual update of the data base descriptions would be sufficient. For these future updates, the data base description would be printed on the turnaround documents and sent to the responsible agencies for updating. Along with the description would be a guide explaining how the turnaround document should be completed, and examples of completed forms to illustrate what type of information should be entered for each data element. The combination of an initial personal interview and subsequent guidelines for updating a description of a data base should provide a convenient and reliable update mechanism for the Chemical Data Bases Directory.

After the turnaround documents have been completed they would be returned to an information specialist for review before the updated information is entered in the CDBD. This review process includes checking the form for completeness and correctness, and determining references for the data element name thesaurus and the subject category thesaurus. The data element name thesaurus would contain a list of standard element names that will be used in the DATA ELEMENTS field of CDBD. If the data elements indicated by the responsible agency on the turnaround document do not match a standard data element name, then either a new data element name would be entered in the thesaurus or the data element name on the turnaround document would be crossreferenced to a data element name in the thesaurus.

The Directory would contain two entries for each data element name. The first would be the standard data element name from the thesaurus, while the second would be the data element name actually used by the particular data base. This would enable a user to search the Directory using a set of standard data element names and then determine the actual data element name to be used when searching the particular data base.

A subject category thesaurus would also be maintained for the CDBD. This thesaurus will consist of a hierarchical scheme of subject category identifiers that will be used to indicate the type of information contained in a particular data base. The thesaurus would be compiled and maintained by an information specialist using documentation and/or personal interview information supplied by the responsible agency for each data base. Based on the thesaurus, entries for the SUBJECT CATEGORY field of the CDBD data base would be assigned by an information specialist prior to the Directory update. This procedure would enable a user to search the Directory for the information content of a data base using a standard predefined set of subject category headings.

Update transactions would be coded from the information on the turnaround document, the standard data element name assignments and the standard subject category entries. These update transactions would be fed into a computer program to update the CDBD data base. Because of the low volume of transactions and the infrequent updating cycle, a batch type update would be sufficient. Since the size of the CDBD data base would be small, the data base records and their associated indexing files could be completely regenerated during each update cycle. This would eliminate the need for a complex file structure and organization to handle additions and deletions. For backup purposes, an extract program would be employed to obtain a copy of the description of each data base in the CDBD and write it onto magnetic tape. The records on tape would be formatted in such a way to enable them to be read by the update program for recovery purposes.

A user would query the Directory using a terminal and receive the results of that query through on-line interactive dialogue between the user and a search and retrieval program. This computer program would support the search and retrieval capabilities discussed in Section 2.3.3.

2.4.2 Plan I - Incorporate CDBD into MEDLARS

The first alternative for implementing the Chemical Data Bases Directory involves incorporating the Inventory of Federal Chemical Information Systems into the MEDLARS System at the National Library of Medicine (NLM). This approach is recommended because the MEDLARS System provides a fully developed search and retrieval, and data organization capability (ELHILL software) that could be utilized to query the CDBD. The ELHILL search and retrieval strategies are also known by a large number of users. The CDBD file organization and structure is compatible with the file organization and structure supported by the MEDLARS System. Another reason for recommending incorporation into MEDLARS is that the system is available on a national timesharing network and already has a large number of users. Any user that has a subscription to any of the other data bases in MEDLARS would also have access to the Chemical Data Bases Directory.

Table 2-2 lists the tasks required to incorporate the CDBD into MEDLARS and a time phased schedule for implementation. Each task is described below:

Task 1 - Data Base Organization

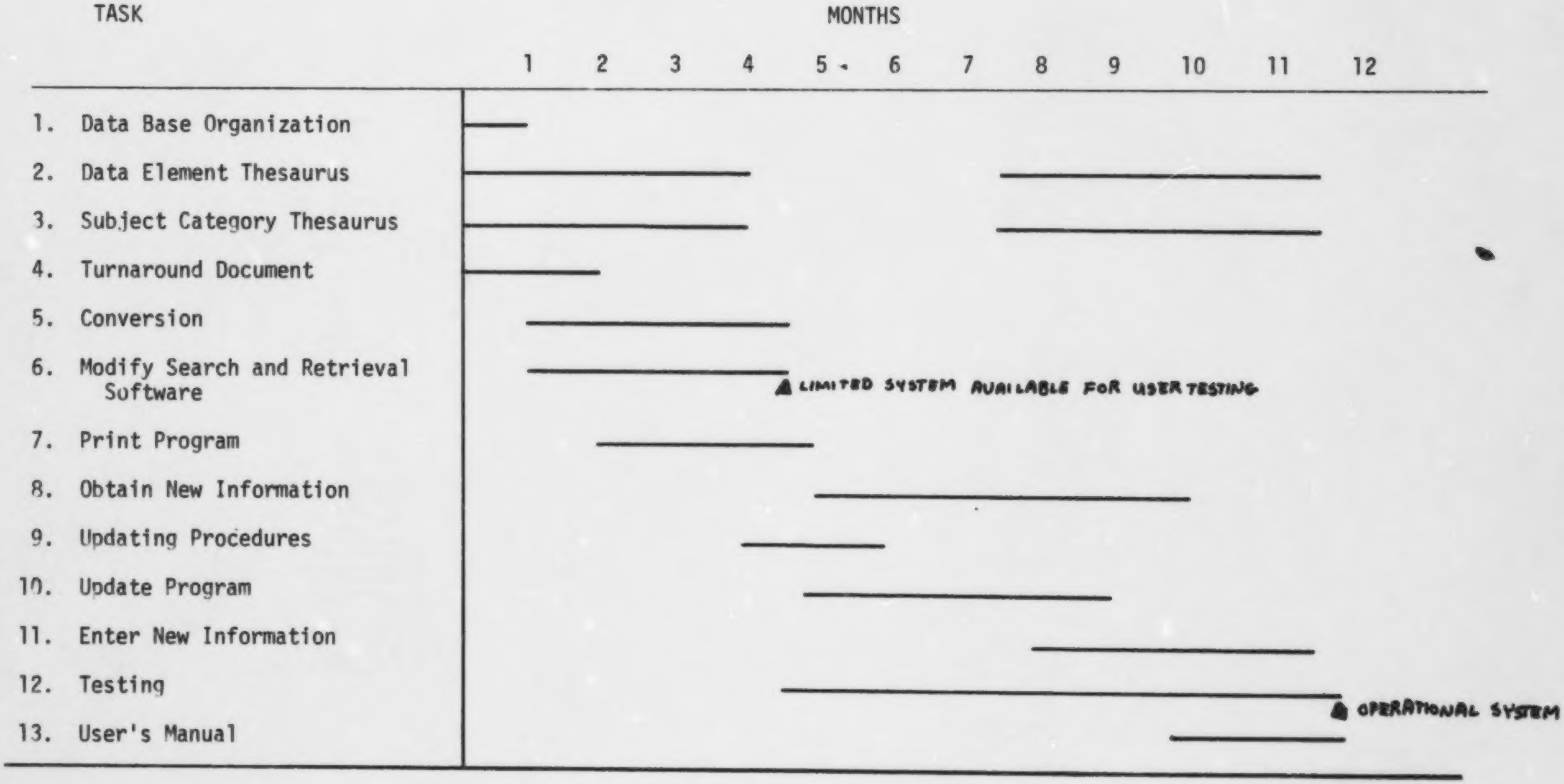
Based on the data elements described in Section 2.3.2, a final decision would have to be made to determine what data elements to actually include in the CDBD. The physical characteristics of each data element would also be defined. The organization and structure of the CDBD would then be defined in a format compatible with the MEDLARS system.

Task 2 - Data Element Thesaurus

A thesaurus of standard data element names would be developed for the CDBD and used when searching the DATA ELEMENTS field of the Directory. A preliminary list of data element names can be obtained from the Inventory of Federal Chemical Information Systems which is located at NLM on a file under the Inquire data base management system (DBMS). Data element names that are similar would be resolved into a single standard data element name. The DATA ELEMENTS field would contain two entries; the standard data element name, and the actual data element name to be used when searching the particular data base. This scheme would provide the user with a uniform search capability.

TABLE 2-2

CDBD PLAN I IMPLEMENTATION TASKS



The data element name thesaurus would be updated as new information about the data bases is obtained from the various agencies. Data element names indicated on the turnaround document that do not match a standard data element name would either be cross-referenced to a standard data element name or designated as a new standard data element name.

Task 3 - Subject Category Thesaurus

A hierarchical index of subject categories describing the information content of the data bases referenced in the Directory would be developed. Since the Directory would be in the MEDLARS Systems, the subject categories could be incorporated into the existing MESH thesaurus, however, a final decision would have to be made whether to utilize MESH or develop an independent thesaurus.

Subject categories will be assigned by an information specialist after analyzing the documentation for each data base and through interviews with the owner of each data base.

Task 4 - Turnaround Document

Based on the data elements and organization developed in Task 1, a turnaround document would be developed to obtain new information and update existing information about the data bases in the Directory. The document would identify all the data elements used to describe a data base, provide space for printing the description of each data element as it exists in the Directory, and provide space for updating the description of each data element.

The turnaround document would be used as a guideline by information specialists when conducting personal interviews, and, for future annual updates, would be used by data base owners to indicate data base description changes. Associated with the turnaround document would be a set of detailed instructions explaining the type and format of information to be entered for each data element. The instructions would include sample turnaround documents that have been filled in.

Task 5 - Conversion

The existing Inventory of Federal Chemical Information Systems on the Inquire data based management system will be used as a basis for the Chemical Data Bases Directory. A conversion program would have to be developed to extract appropriate information from the Inventory and construct the appropriate Header, Index, and Posting files in the MEDLARS system. New data elements for which no information exists in the Inventory would contain blanks in the MEDLARS CDBD.

Task 6 - Modify Search and Retrieval Software

The ELHILL software would be utilized to perform the search and retrieval functions for the CDBD. The appropriate modifications would have to be made to the ELHILL software to identify the various data sets comprising the CDBD and to identify the searching and printing characteristics of each data element in the CDBD. At the completion of Tasks 5 and 6, the Directory could be accessed in a test mode on a limited basis. At that time, the information converted from the Inventory would be available for querying by the user. This will establish an immediate testing environment so that required changes could be identified early and incorporated into the operational CDBD.

Task 7 - Print Program

Based on the turnaround document developed in Task 4, a computer program would be developed to print a description of each data base that exists in the CDBD. The printed turnaround documents will provide the update mechanism for the Directory.

Task 8 - Obtain New Information

Since the Directory that was established from the Inventory of Federal Chemical Information Systems would not contain information on new data elements defined in Task 1, additional data would have to be obtained from the agencies responsible for the various data bases. It is recommended that personal interviews be conducted in order to obtain the most complete and accurate information required for the operational version of the Directory. The turnaround documents would be used as a guideline when conducting the interviews. These personal interviews would also provide a means of explaining the purpose of the Directory and the various data elements that describe the data base. Establishing a good working relationship with the agencies will be important for obtaining future updates to the Directory. Since the number of data bases in the Directory is large, they should be ranked in order of importance when developing an interview schedule.

Task 9 - Updating Procedures

Procedures would be developed to analyze the turnaround documents for completeness and accuracy, and for entering the information into the Directory. The format of update transactions would be defined as well as the procedures for creating the transactions from the turnaround documents. Procedures for determining the standard data element names and subject categories for each data base as well as updating the data element and subject category thesaurus would be defined.

Task 10 - Update Program

A batch type computer program would be developed to apply the update transactions to the Directory. The program would construct the Header, Index and Posting files according to the MEDLARS format. Because the size of the Directory data base is small, a new copy would be recreated during each update cycle. For backup purposes the update program would create a copy of the Directory Header records each time it is run. The format of the backup records would be such that the Directory could be reconstructed by processing them through the update program.

Task 11 - Enter New Information

As new information is obtained from the various agencies, the turnaround documents would be processed and applied to the Directory. Appropriate standard data element names and subject categories would be assigned to each data base. The data element and subject category thesaurus would be updated accordingly. Update transactions would be created from each turnaround document and processed by the update program.

Task 12 - Testing

At the conclusion of Tasks 5 and 6, a limited version of the Directory would be available for testing. Making the Directory available at an early date would enable users to identify problems or changes that may require system modifications. The long lead time before full operation would permit successful implementation of required modifications. System testing would continue until the Directory becomes fully operational at the end of the twelve month implementation period.

Task 13 - Users Manual

An explanation of how to use the Directory would be developed and incorporated into the MEDLARS Users Manual. It would describe the information content of the Directory, list and explain each data element, and illustrate examples of typical searches and retrievals available to the users.

Under this plan, a fully operational CDBD could be incorporated into the MEDLARS system within one year. The level of effort to implement Plan I is estimated at 10 months for a senior systems analyst, 12 months for a senior information specialist and 6 months for a junior information specialist. The cost to perform the 13 tasks described above is estimated at $125,000.

Once the CDBD becomes operational, it is recommended that the data base descriptions be updated on an annual basis. The updating process would include printing the turnaround documents, distributing them to the various agencies, editing the updated turnaround documents, updating the data element and subject category thesaurus, creating update transactions, and applying the transactions to the Directory. Part of the update procedure would be to identify additional data bases to be included in the Directory. It is estimated that the annual update procedure would require .5 months for a systems analyst and 2 months for an information specialist.

User support and training could be included in the MEDLARS support team and should require no more than a month of effort per year.

2.4.3 Plan II - Incorporate CDBD into the Existing Inventory of Federal Chemical Information Systems

The second alternative for implementing the Chemical Data Bases Directory involves modifying and enhancing the existing Inventory of Federal Chemical Information Systems that is processed using the Inquire data base management system (DBMD) at NLM. This alternative is proposed because the Inventory provides an existing foundation for the CDBD, the Inquire DBMS has an on-line updating capability, the computer at NLM is connected to a national time-sharing network, and the on-line query and reporting capabilities of the Inquire DBMS could be used for adhoc user requirements.

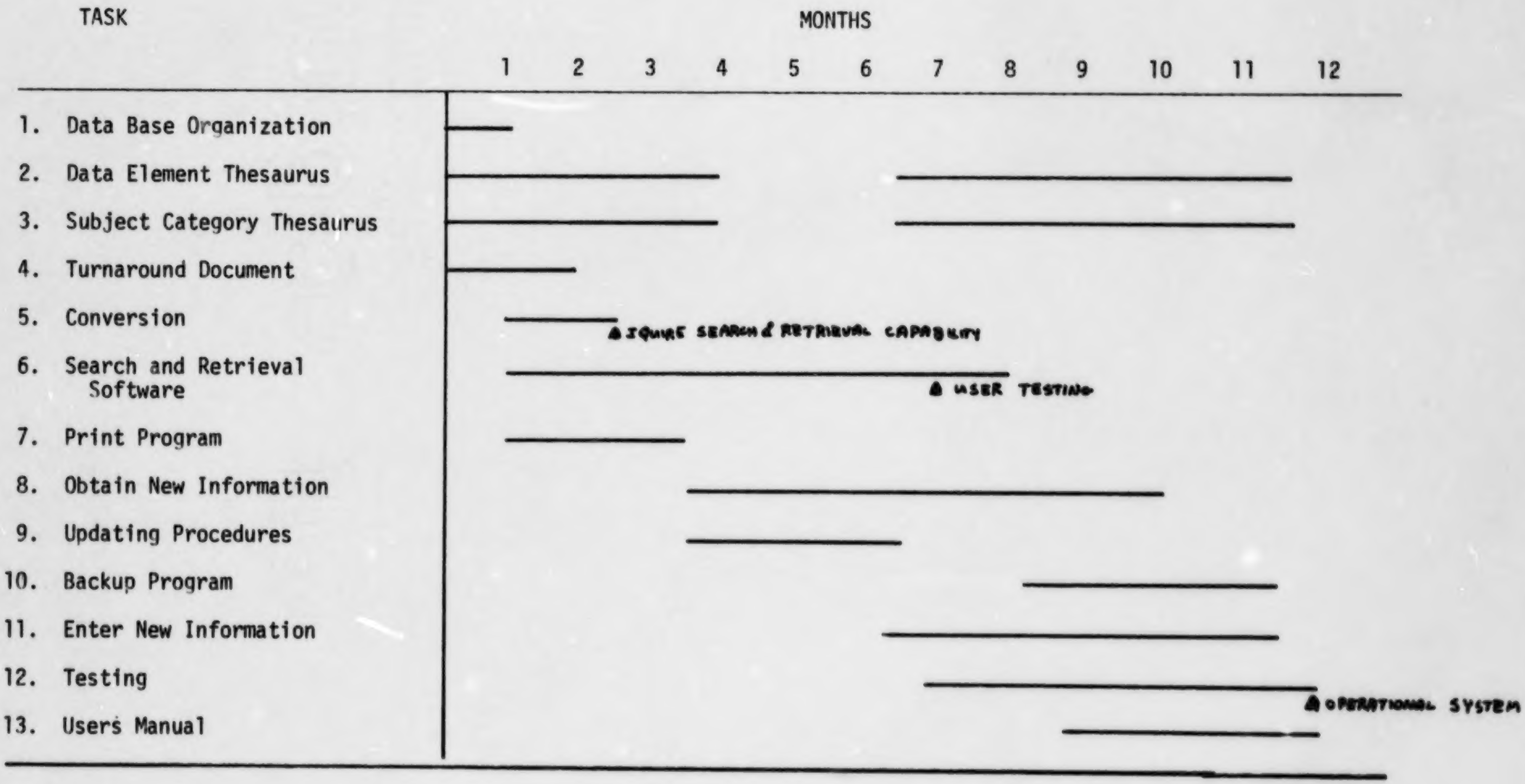
Table 2-3 lists the tasks required to incorporate the CDBD into the existing Inventory and a time phased schedule for implementation. Each task is described below:

Task 1 - Data Base Organization

Based on the data elements described in Section 2.3.2, a final decision would have to be made to determine what data elements to actually include in the CDBD. The physical characteristics of each data element would also be defined. A data base description would then be defined to the Inquire DBMS, using as much as possible of the data base description of the current Inventory of Federal Chemical Information Systems.

TABLE 2-3

CDBD PLAN II IMPLEMENTATION TASKS



Task 2 - Data Element Thesaurus

This task is identical to Task 2 described in Section 2.4.2.

Task 3 - Subject Category Thesaurus

This task is identical to Task 3 described in Section 2.4.2, except that the thesaurus could not be included in the existing MESH thesaurus used by the MEDLARS system.

Task 4 - Turnaround Document

This task is identical to Task 4 described in Section 2.4.2.

Task 5 - Conversion

A special program would be developed to extract appropriate information from the current Inventory of Federal Chemical Information Systems and create the initial Directory as another Inquire data base using the data base description defined in Task 1.

Task 6 - Search and Retrieval Software

The Inquire DBMS provides an on-line query and reporting capability as part of its system package. However, it is not user prompted, somewhat difficult to use, and requires at least a three day training session. It is therefore recommended that a user prompted search and retrieval computer program be developed to interface between the user and the Directory. The program would support the search and retrieval capabilities discussed in Section 2.3.3.

Task 7 - Print Program

This task is identical to Task described in Section 2.4.2.

Task 8 - Obtain New Information

This task is identical to Task 8 described in Section 2.4.2.

Task 9 - Updating Procedures

Procedures would be developed to analyze the turnaround documents for completeness and accuracy. Procedures for determining the standard data element names and subject categories for each data base as well as updating the data element and subject category thesaurus would be defined. Since the Inquire DBMS provides an on-line updating capability as part of its system package, it is recommended that this feature be utilized. By using this feature, information from the turnaround documents could be directly applied to the CDBD, instead of creating transactions that would be used by an update program to apply the data to the CDBD.

Task 10 - Backup Program

To provide a backup and recovery capability for the CDBD, a backup program would be developed and run after each update cycle. The format of the backup records will be such that the CDBD could be restored using an Inquire DBMS utility to create a data base. The backup file could also be used to provide a copy of the Directory for another computer or installation.

Task 11 - Enter New Information

This task is identical to Task 11 described in Section 2.4.2, except that the CDBD will be updated on-line instead of creating transactions and executing an update program.

Task 12 - Testing

At the conclusion of Task 6, the Directory will be available for user testing in order to provide feedback on the system design. During this five month test period before the Directory becomes fully operational, problems or changes that may require system modifications would be identified and corrected.

Task 13 - Users Manual

An explanation of how to use the Directory would be developed and made available to users. It would describe the information content of the Directory, list and explain each data element, explain the search and retrieval strategies, and illustrate examples of typical searches and retrievals available to the user.

A fully operation CDBD could be incorporated into the Inventory of Federal Chemical Information's System within one year. The level of effort to implement Plan II is estimated at 12 months for a senior systems analyst, 5 months for a programmer, 12 months for a senior information specialist and 6 months for a junior information specialist.

The cost to perform the 13 tasks described above is estimated at $160,000. Once the CDBD becomes operational, it is recommended that the data base descriptions be updated on an annual basis. The updating process would include printing the turnaround documents, distributing them to the various agencies, editing the updated turnaround documents, updating the data element and subject category thesaurus, and updating the Directory with the information on the turnaround documents. Part of the update procedure will be to identify additional data bases to be included in the Directory. It is estimated that the annual update procedure would require .5 months for a systems analyst and 2 months for an information specialist.

A user support and training group would have to be developed. This function is estimated to require no more than one month of effort per year and could be incorporated into the annual update function.

2.4.4 Plan III - Develop the CDBD Using Custom Written Software

The third alternative for implementing the Chemical Data Bases Directory involves designing a system using no proprietary software packages. This means that all file structures and computer programs for search and retrieval, updating, and printing would have to be custom designed and written. By using this approach to implement the CDBD, a great deal of transportability would be built into the system. This would enable the CDBD to be transferred to another computer or another installation with a minimum amount of change.

Table 2-4 lists the task required to complete Plan III and a time phased schedule for implementation. Each task is described below:

Task 1 - Data Base Organization

Based on the data elements described in Section 2.3.2, a final decision would be made to determine what data elements to actually include in the CDBD. The physical characteristics of each data element would also be defined. The organization and structure of the CDBD would not incorporate any computer dependent characteristics. A simple indexing scheme should be sufficient.

Task 2 - Data Element Thesaurus

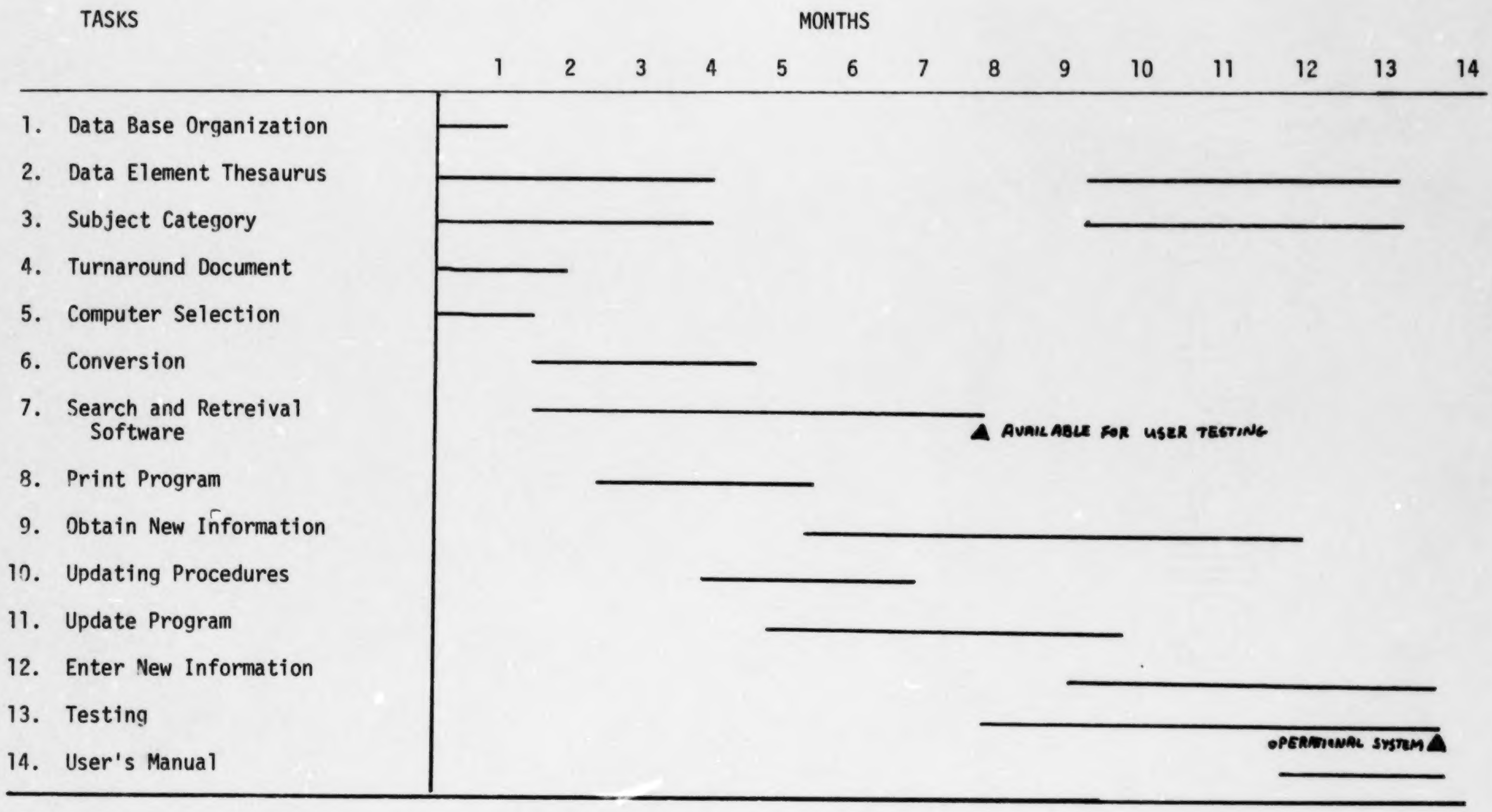
This task is identical to Task 2 described in Section 2.4.2.

Task 3 - Subject Category Thesaurus

This task is identical to Task 3 described in Section 2.4.2, except that the thesaurus could not be included in the existing MESH thesaurus used by the MEDLARS system.

TABLE 2-4

CDBD PLAN III IMPLEMENTATION TASKS



Task 4 - Turnaround Document

This task is identical to Task 4 described in Section 2.4.2.

Task 5 - Computer Selection

A computer and installation would have to be selected to process the CDBD. A large mainframe timesharing computer or a minicomputer would be adequate to accommodate the CDBD. The installation chosen would have to provide an on-line networking capability to allow a large number of users access to the Directory.

Task 6 - Conversion

A special program would be developed to extract appropriate information from the current Inventory of Federal Chemical Information Systems and create the initial Directory according to the data base organization and structure defined in Task 1.

Task 7 - Search and Retrieval Software

A user prompted search and retrieval computer program would be developed to interface between the user and the Directory. The program would support the search and retrieval capabilities discussed in Section 2.3.e. At the conclusion of this task, the Directory will be available on a limited basis for user testing.

Task 8 - Print Program

This task is identical to Task 7 discussed in Section 2.4.2.

Task 9 - Obtain New Information

This task is identical to Task 8 discussed in Section 2.4.2.

Task 10 - Updating Procedures

This task is identical to Task 9 discussed in Section 2.4.2.

Task 11 - Update Program

A batch type program would be developed to apply the update transactions to the Directory. The program would construct the data and index files defined in Task 1. Because the size of the Directory data base would be small, a new copy would be recreated during each update cycle. For backup purposes, the update program would create a copy of the Directory each time it is run. The format of the backup records would be such that the Directory can be reconstructed by processing them through the update program.

Task 12 - Enter New Information

This task is identical to Task 11 discussed in Section 2.4.2.

Task 13 - Testing

This task is identical to Task 12 discussed in Section 2.4.3., except that the testing period in this case would be 6 months.

Task 14 - Users Manual

This task is identical to Task 13 discussed in Section 2.4.3.

Plan III could be fully operational within 14 months. The level of effort to implement Plan III is estimated at 14 months for a senior systems analyst, 10 months for a programmer, 12 months for a senior information specialist, and 6 months for a junior information specialist. The cost to perform the 14 tasks described above is estimated to be $190,000.

The annual updating and user support and training requirements are the same as those discussed in Section 2.4.3.

2.4.5 Advantages and Disadvantages of Each Plan

A comparison of the three implementation alternatives is presented in order to identify their strengths and weaknesses. Table 2-5 is a summary of the advantages and disadvantages of each Plan.

Plan I involves incorporating the CDBD into MEDLARS. The main advantages to this approach are that MEDLARS is already connected to a national timsharing network (Tymeshare), and a fully operational search and retrieval software package (ELHILL) could be used to access the Directory. The ELHILL software is easy to learn and is already known by many people. CDBD could easily be incorporated into the MEDLARS user support and training staff. Once the user is connected to MEDLARS, other data bases important to chemical information could be linked to directly.

One disadvantage to this plan is that a user must meet certain requirements before a subscription to MEDLARS can be obtained. Another disadvantage is that search and retrieval and data base design are bound by the design and format requirements of the ELHILL software and MEDLARS data base format.

TABLE 2-5

A COMPARISON OF CDBD IMPLEMENTATION PLANS

PLAN	ADVANTAGES	DISADVANTAGES
I. Incorporate CDBD into MEDLARS	· Communications network already exists · Existing search and retrieval capability · Search & retrieval easy to learn and already known by many people · Direct link to CHEMLINE, TOXLINE, CANCERLINE, MEDLINE · Existing user support and training staff	· Users must meet certain requirements for MEDLARS subscription · Search and retrieval, and data base must be adapted to MEDLARS Format
II. Incorporate CDBD in the Inventory of Federal Chemical Information Systems	· Existing on-line updating capability · Ad hoc user query and report capability · NLM computer already connected to national timesharing network	· User prompted search & retrieval capability will have to be developed · Data base design must be adapted to Inquire DBMS format
III. Develop CDBD using custom written software	· Easily transferred to another computer or installation · Search and retrieval, updating, and data base design do not have to be adapted to existing software	· Longer implementation time required · Greater development cost · Communications network may have to be developed · All software must be designed, written and tested.

Plan II involves incorporating the CDBD into the existing Inventory of Federal Chemical Information System. One advantage to this plan is that the Inquire DBMS, which is used to process the Inventory, provides an on-line updating capability that could be used to enter information into the Directory. The Inquire DBMS also provides an on-line query and reporting capability that could be used for ad hoc queries. However, a specialist would be required to conduct these queries since the quering and reporting requires an in depth knowledge of the Inquire DBMS facilities. Since the Inventory exists on the NLM computer, which is already connected to a national timesharing network, access by many people would not pose a problem.

A disadvantage to Plan II is that a generalized user prompted search and retrieval program would have to be written. As discussed above, the Inquire DBMS capabilities in this area would not be adequate for the general user. Search and retrieval programs usually require a greater investment in manpower time and costs, and represents the biggest difference between Plan I and Plan II. Another disadvantage is that the data base design must conform to the requirements of the Inquire DBMS.

Plan III involves developing the CDBD using custom written software. Using this approach the CDBD could be designed so that it could be transferred to another computer or installation with a minimum amount of difficulty. The search and retrieval, updating, and data base organization could be designed to meet user needs and not be bound or adapted to the requirements of any software package.

A disadvantage to this approach is that custom designed software takes a longer time to develop, usually costs more, and requires an extensive amount of testing before becoming operational.

3.0 CHEMICAL STRUCTURE/NOMENCLATURE SYSTEM

3.1 Introduction

In this section, system requirements and implementation plans are discussed for developing an on-line Chemical Structure/Nomenclature System (CSNS). The CSNS, which will be searchable by a user, will contain identifying information about chemical substances. The identifying information will include chemical names, molecular formula, CAS registry number, Wiswesser line notation, structural features, etc. It will be a comprehensive system in that it will include as much identifying information as possible about as many chemical substances as possible. It is estimated that there will be as many as 500,000 chemical substances in the system. Associated with each chemical substance will be a locator that identifies the data bases that contain other information about the substance. Data bases referenced by CSNS will include those identified in the Chemical Data Bases Directory (CDBD) described in Section 2.3.

CSNS would act as a starting point for a user attempting to identify chemical substances. Since CSNS is a comprehensive system, it will provide the most likely chance of obtaining the desired matches. Once the desired matches are found the user will display the locator list for each substance to identify the data bases containing other information about the chemical substance. This procedure will eliminate the need to search each data base independently in order to identify a substance.

If the locator indicates a data base with which the user is not familiar, the CDBD will be accessed to determine the type of information contained in the data base and the method for searching the data base.

Advantages exist in accessing all chemical data bases using a single universal data element, such as the CAS registry number.

A user could first access CSNS to identify a chemical substance and obtain the associated CAS registry number; once the number was obtained, the user could then access other data bases using only the CAS registry number. This procedure will save storage space and access time on the other data bases.

3.1.1 Background

A comprehensive CSNS was identified as one of the core components of the Chemical Substances Information Network (CSIN) in a previous study completed by the Mitre Corporation. This study of user requirements recommended that CSNS act as a focal point for users desiring information about chemical identification. If will be publicly available and serve users involved in the study and in the regulation of chemicals.

The CSNS is the second critical element in the CSIN. User requirements for CSNS do not depend on the various TSCA reporting requirements for the three development options discussed in the previous study. Therefore, implementation of CSNS can be accomplished independently of other components in CSIN.

3.1.2 Approach

There currently exist a number of systems that could be classified as chemical structure and nomenclature systems. The Mitre report identified two systems that met most of the requirements of CSNS; CHEMLINE and SSS/CIS. The report also mentioned that certain features were available in CIDS. A fourth system, applicable to CSNS, DR&D CIS, was identified by Sigma Data. An in-depth analysis of these four systems was conducted to determine their capabilities for satisfying the proposed Chemical Structure/Nomenclature System (CSNS).

Based on the recommendations of the Mitre Report, a comprehensive set of system requirements was developed for CSNS. The requirements were divided into nine categories: subject matter and scope, data elements, search and retrieval capabilities, user access, user support, maintenance and update, file organization and structure, hardware, and software.

A subjective quantitative comparison was developed to determine how well the four systems met the system requirements of CSNS.

Based on the comparison, two systems, CHEMLINE and SSS/CIS, were selected as possible candidates for the Chemical Structure/Nomenclature System. Three alternative implementation plans were developed which included enhancing either one or both of the systems. Each plan discussed the tasks required to integrate the CSNS into the Chemical Substances Information Network and considered manpower, costs, and time frames associated with the development. A comparison of each plan was presented by discussing their advantages and disadvantages.

3.2 Existing Capabilities

This section describes four operational systems that are representative of chemical structure and nomenclature systems. An analysis of their capabilities will determine how effectively they meet the system requirements of CSNS. Information for the analysis was obtained from interviews with administrative and technical personnel associated with the system, user's manuals, live demonstrations, and actual user access by Sigma Data personnel. The description of each system is divided into six parts:

A. Information Content
B. Organization and Structure
C. Retrieval Process
D. Hardware and Software
E. User Support
F. Future Plans

3.2.1 CHEMLINE - Chemical Information On-Line

A. Information Content

CHEMLINE is an on-line chemical dictionary data base that provides a mechanism whereby chemical substances can be searched and retrieved using chemical names, numbers, and structural criteria. It is a subset of the MEDLARS system which is maintained at the National Library of Medicine (NLM).

CHEMLINE contains records for all chemical substances cited in the TOXLINE bibliographic files, another subset of MEDLARS. It also contains some chemical substances cited in CANCERLINE and will contain the TSCA candidate list in July 1977. Currently there are about 100,000 unique chemical substances in the data base. A user searches CHEMLINE before searching TOXLINE in order to determine the CAS registry number and various names by which a chemical substance may be identified in TOXLINE.

Table 3-1 is a list of data elements contained in CHEMLINE including an indication if it is searchable and/or printable.

TABLE 3-1

CHEMLINE DATA ELEMENTS

NAME	SEARCHABLE	PRINTABLE
CAS Registry Number	X	X
Alternate Registry Number	X	X
Molecular Formula (Hill)	X	X
Molecular Formula Fragments	X	
CAS Preferred Name	X	X
Other Chemical Names and Synonyms	X	X
Name Fragments	X	
Wiswesser Line Notation (WLN)	X	X
Number of Rings	X	X
Ring Size	X	X
Ring Size Fragments	X	
Ring Elemental Analysis	X	X
Ring Elemental Analysis Fragments	X	
Component Line Formula	X	X
Component Line Formula Fragments	X	
Locator	X	X

B. Organization and Structure

The CHEMLINE data base consists of three internal files:

- Index File
- Posting File
- Header or Data File

The Header file consists of all the unit records which contain the information for a particular chemical substance. During the data base update procedure, each new Header file record is identified by a unique Computer Assigned Number (CAN). This CAN number is four bytes long and contains disk location information to provide rapid access to the Header file during the on-line retrieval process.

The Index file consists of an entry for each unique occurrence of a data element designated as a searchable term in the unit record. The Index file acts as an inverted list to the search terms of the unit records. The entries are arranged in one continuous alphabetic listing, A-Z, followed by decimals 0-9. Each entry also consists of a two-part number. The first part of the number is an address or location of entries in the Posting file, for the particular Index file entry. The second part of the number indicates how many entries there are in the Posting file for the particular Index file entry.

The Posting file consists of strings of CAN numbers corresponding to the unit records associated with each search term entry in the Index file. It provides the link between the Index file and the Header file.

This type of structure provides an efficient method of answering and storing the intermediate results of on-line queries. Intermediate or final results of searches are stored as lists of CAN numbers retrieved from the Posting file.

The CHEMLINE data base is currently updated two or three times a year. All new CAS registry numbers and/or names that have been added to the TOXLINE file since the last update are tape extracted and sent to the Oak Ridge National Laboratory (ORNL) to be processed. ORNL, which has a copy of the Chemical Abstracts Service registry system, matches the extract tape and produces a new tape containing CAS number, CAS names, molecular formula, and ring information. This tape is returned to NLM, where WLN and locator information is added. The resultant file is then used to update the Header file, where new records are added to the end of the current Header file and new CAN numbers are generated. The Index and Posting files are totally recreated after each update to facilitate minimum search time.

C. Retrieval Process

Access to the CHEMLINE data base is available through the Tymeshare network or through a direct dial-up to NLM. Since CHEMLINE is available through Tymeshare, any terminal supported by this network can be used to access CHEMLINE information. Currently these include:

- TWX
- Teletype
- Teletype Compatible
- IBM 2741 Compatible
- CRT

CHEMLINE is used mainly to search for the existence of a certain chemical substance and determine its CAS registry number and its various names. Once the numbers and/or names are obtained they can be used to search TOXLINE or other relevant data bases for the desired information about a substance.

The user searches the data base using the retrieval commands of the ELHILL software. The following data items may be combined in various ways using logical connectives to produce the desired search strategy:

CAS REGISTRY NUMBER - 9-digit number in hyphenated format without leading zeros, e.g., 468-86-2.

ALTERNATE CAS REGISTRY NUMBER - CAS registry numbers which are cross-referenced to current or superceded registry numbers.

MOLECULAR FORMULA - The molecular formula is expressed in the Hill convention. Salts are designated in the "dot-disconnect" fashion, i.e., C6-H8-07.3NA. The "NEIGHBOR" command can be used to obtain all of the salts for a substance when the user is in doubt about the particular salt.

MOLECULAR FORMULA FRAGMENTS - The user can search for any combination of individual elements except for hydrogen. This system does not support ranges of occurrences of a particular element.

CAS PREFERRED NAME - This is the Type 1 systematic name produced by CAS and includes both the 8th and 9th collective index. Only the first 39 characters of the name are searchable.

OTHER CHEMICAL NAMES AND SYNONYMS - These include the Type 2 and Type 3 names including all synonyms.

NAME FRAGMENTS - This data item consists of all meaningful parts of the chemical names, each of which is independently searchable. A fragment is created for each break on all closures, parentheses, hyphens, colons, and blanks.

WISWESSER LINE NOTATION - Currently only about 10% of the CHEMLINE records contain a WLN.

NUMBER OF RINGS - Indicates the number of rings in the structure of the chemical substance.

RING SIZE - Indicates the size of each ring in the ring system.

RING SIZE FRAGMENT - The user can search for a specified ring size imbedded in a ring system.

RING ELEMENTAL ANALYSIS - The user can search for the elements that make up the individual rings in a ring system. A hyphen is used to separate each ring.

ELEMENTAL FRAGMENTS - Breaks the ring elemental analysis on each hyphen so that each individual ring is independently searchable.

COMPONENT LINE FORMULA - Denotes the atom sequence in all ring components that contain at least two non-carbon atoms.

COMPONENT LINE FORMULA FRAGMENTS - Breaks the component line formula on each ring so that individual rings are independently searchable.

Each successful seach strategy or query results in the creation of a temporary file that contains a list of the CAN numbers that satisfied the search. This temporary file can be further refined or modified by combining it with other temporary files or search statments using logical connectives (AND, OR, NOT) to produce another temporary file. The user is allowed to create a maximum of 30 temporary files for each session. These temporary files are deleted at the end of the session.

The user has the option of printing the information satisfying a search query by specifying the appropriate temporary file. The list of CAN numbers in the temporary file are used to locate the unit records in the Header file from which the designated information is displayed on the terminal. By printing the information in the locator data element, the user can determine which files contain information about a particular chemical substance.

D. Hardware and Software

The CHEMLINE data base physically exists at NLM and is operational on an IBM 370/158 under VS with 4M bytes of core storage. The files are stored on IBM 3350 disk packs and use about 400M bytes of storage.

On-line retrieval sessions for CHEMLINE are processed using the ELHILL-3 software package developed by Systems Development Corporation (SDC). The current version is the result of modifications made to ORBIT, ELHILL-1 and ELHILL-2. The software is modularly designed to minimize the modification effort. ELHILL-3 is programmed mostly in PL/1 with some assembler language (ALC) routines to provide a more efficient input-output capability. The software contains an overlay structure which enables it to operate in a maximum of about 300K bytes of core. The assembler language routines, written expressly for an IBM computer, pose the most significant restrictions to converting the software to another machine or system. Currently two copies of the ELHILL software are operating simultaneously to serve a maximum of 120 users. However, a more realistic maximum due to response time is 75-80 users. There are no unusual terminal use patterns or non-standard terminal features.

Since the ELHILL software is mostly table-driven or specification-driven, additions or modifications to CHEMLINE are relatively easy. For instance, to add another data element to the file, a new entry in the specification table for CHEMLINE is coded and cataloged into the ELHILL software system. The update program, which is modular, is modified to perform any required editing on the new data element. The ELHILL retrieval program, since it utilizes table look-up procedures, does not require any modification.

The file organization and assembler language routines performing input-output processing are tailored to an IBM 370 computer. Converting to a machine other than an IBM 360/370 would require extensive modification.

E. User Support

Users of CHEMLINE are provided with a comprehensive and easily readable reference manual produced by NLM. The reference manual describes the procedures necessary to access the system together with the search and retrieval capabilities available to the user. The data element information content is described in detail along with examples of typical search strategies the user might employ.

NLM also conducts formal training sessions for users with varying levels of expertise. The user can obtain general and technical assistance by calling a specified telephone number. Questions on system and communication problems and program interpretation are answered by support staff.

F. Future Plans

NLM is currently in the process of making the TSCA inventory list of about 36,000 chemical substances available to users. The list should be available in July 1977.

NLM is also testing an 8,000 chemical substance subset of CHEMLINE to determine the feasibility of implementing a more comprehensive name fragmentation system. Currently name fragments are broken on closures, hyphens, parentheses, commas, and blanks. Chemical Abstracts Service has developed a more comprehensive fragmentation scheme that splits suffixes and prefixes. For instance, the currently searchable name fragment data item, monohydrochloride, would be fragmented into three searchable data items: mono, hydro, and chloride. If the testing proves favorable the new name fragmentation scheme will be incorporated into CHEMLINE.

In July 1977 the number of unique chemical substances in CHEMLINE will increase to 225,000 from the current 100,000. This is due to a special extract by CAS to obtain all chemical substances that were cited in the text but not in the abstract of TOXLINE citations.

NLM is also considering a feasibility study to obtain CAS connection tables and the structural keys (pre-formatted screens) for all chemical substances. A user could then search CHEMLINE using these predefined structural keys and could display a two dimensional structure diagram of the chemical substance using the stored connection table. The user would not be able to conduct an atom-by-atom search using the connection tables.

3.2.2 SSS/CIS - Substructure Search of the Chemical Information System

A. Information Content

SSS is part of the Chemical Information System (CIS) developed by NIH/EPA. The data base contains chemical structure information about substances cited in various data bases. Currently, these data bases include:

- Mass Spectrometry
- Oil and Hazardous Materials
- Carbon 13 NMR
- TSCA Inventory List
- EPA Pesticides

The system allows the user to define a chemical structure or substructure and then to search for occurrences of the structure or substructure in the various data bases of the NIH/EPA CIS. Currently SSS contains information on nearly 70,000 chemical substances; other data bases will be added to the system during 1977 to bring the total to around 250,000 substances.

The user accesses the system using terminals in a conversational interactive mode through the Tymshare Network. Table 3-2 is a list of data elements contained in SSS.

B. Organization and Structure

Currently the SSS data base consists of a set of files for each of the five data bases that make up the SSS system:

- Mass Spectrometry
- Oil and Hazardous Materials
- Carbon 13 NMA
- TSCA Candidate List
- Pesticides

Each set of files consists of the following data sets:

TABLE 3-2

SSS DATA ELEMENTS

NAME	SEARCHABLE	PRINTABLE
CAS Registry Number	X	X
CAS Chemical Names		X
Molecular Formula	X	X
Molecular Formula Fragments	X	
Two Dimensional Structure Representation	X	X
Structural Feature Codes	X	
Nucleus Probe	X	
Molecular Weight	X	
Total Number of Atoms	X	
Mass Spectrometry	X	
Atom Centered Fragments	X	
Ring Count/Population	X	
TSCA Inventory Number		X

- Structure File
- Index to Structure File
- Nomenclature File
- Index to Nomenclature File
- Search Files (35-40)

The Structure file contains the CAS connection tables for all substances. It is used to do an atom-by-atom substructure search and to display the two dimensional chemical structure. The Nomenclature file contains the chemical name, molecular formula, etc. The Indexes to the Structure and Nomenclature files are inverted lists of CAS registry numbers that point to a particular structure or nomenclature record.

The Search files are inverted lists of values of various data elements that are used to satisfy a particular search query. Associated with each value of a data element is a list of CAS registry numbers that satisfy the particular value. By splitting the various searchable data elements into a number of files, the amount of search time required to satisfy a particular query is reduced.

The CAS registry number is used to link the various files. Results of searches are stored as temporary files which consist of lists of CAS registry numbers.

The five searchable data bases are currently updated independently and on an as-needed basis. Fein-Marquart Associates is in charge of maintaining and updating SSS.

Every time a file is updated a new set of SSS files is created. Tapes are received from CAS containing structure and nomenclature information; these tapes are used to create the Structure, Nomenclature and related Index files. The CAS structure information is then processed to create the various Search files.

C. Retrieval Process

Since SSS is available through Tymshare Network, any terminal supported by this network can be used to access SSS information. Currently these include:

- TWX
- Teletype
- Teletype Compatible
- IBM 2741 Compatible
- CRT

A maximum of 60 users can be served simultaneously; however, a more realistic maximum due to response time is 35-40 users. There are no unusual terminal use patterns or non-standard terminal features. SSS optionally supports certain sophisticated terminals with vectoring capabilities to provide a more meaningful structure representation.

SSS is used to search a specified data base for chemical substances satisfying a particular combination of structural and non-structural requirements. Initially, the user must indicate which of the five representative files is to be searched. Currently, there is no provision for combining several files into a single search session; all files must be searched independently.

Two methods are available to the user to search the data base: by a chemical structure, or by chemical identification characteristics. The chemical structure method consists of "drawing" a two dimensional pictorial representation of the structure of a substance. Various terminal commands are available to "draw" the structure, including specifying rings, chains, branches, atoms, bonds, links, and nuclei. At any time during construction, the user can display the structure to ensure the desired representation.

As an alternative to "drawing" a structure, the user can specify that the query structure be that of a known substance by specifying the CAS registry number. The user then has the capability to use the structure generation commands to modify the structure as desired for the actual query.

Once a structure has been defined, the user can search the data base against the structure in three ways:

- Fragment Probe
- Nucleus Probe
- Molecular Formula

The fragment probe is designed to search for all unique atom-centered fragments, which consists of an atom and all immediate neighbors. The nucleus probe will search the data base for each nucleus within the query structure. The molecular formula can be searched as a total molecular formula, partial formula, or ranged partial formula.

As an alternative to searching the data base against a query structure, the user can specify certain substance identifiers or characteristics. These include:

CAS NUMBER - Chemical Abstract Services unique identifying number.

MOLECULAR FORMULA - Described above.

MOLECULAR WEIGHT - A specified weight or a range of weights. Hydrogen atoms are not included.

TOTAL ATOM COUNT - A specified count or range. Hydrogen atoms are not included.

RING COUNT/POPULATION - Total number of rings, ring sizes.

STRUCTURAL FEATURE CODES - Predefined chemical structure characteristics that are identified by a unique key. These keys are identical to the keys used in CIDS (Chemical Information and Data System) at Edgewood Arsenal.

The results of a successful search consists of a list of CAS registry numbers (stored in a temporary file) that match the search criteria. Each temporary file is assigned a unique number. A user is allowed to create as many as 30 temporary files in a session. Temporary files can be combined using logical connectives (AND, OR and NOT) to create new temporary files.

For any temporary file the user has the capability of performing a substructure search. This capability enables the user to ensure that responding substances exactly match all structural requirements of the query structure on an atom-by-atom, bond-by-bond basis. Since this is a relatively time-consuming process costs should be considered before searching temporary files containing more than 100 substances.

SSS provides the capability of saving a query structure for re-use during a future terminal session. This can be useful for complicated structures because the structure definition commands require a considerable amount of terminal time to input.

The user can display the results of a query either on the terminal or off-line. The following items are printed:

- CAS Registry Number
- Moledular Formula
- CAS Names
- Two Dimensional Pictorial Representation of the structure
- TSCA Inventory Number (when applicable)

The user can switch from one file to another at any time during the query session.

Chemical names can not be searched in any way on SSS.

D. Hardware and Software

The SSS data base physically exists at the Tymshare center in Houston, Texas, on a DEC/PDP10. The files are stored on disk packs that use about 35M bytes of storage. All of the data base preprocessing is done on an IBM 370/168.

On-line retrieval sessions for SSS are processed using a software package originally developed by NIH and later modified by Fein-Marquart Associates under contract to EPA. SSS is programmed mainly in ANSI Fortran with some assembler language routines to provide a more efficient I/O capability. The software contains an overlay structure which enables it to operate in a maximum of 24K words of core storage. The assembler language routines, written expressly for a DEC/PDP 10 computer, pose the most significant restrictions to converting the software to another machine or system.

SSS software is modularly designed to facilitate easy modification. To add a searchable data element to the SSS system, a new Search file must be created. This is accomplished by adding the appropriate module to the update program. The SSS retrieval program requires a new module to process any new commands needed to search for the new element.

Since each of the five searchable data bases contains its own unique set of files for processing, incorporating an additional data base is fairly easy. Because each file in the set has the same type of file organization, the file organization and assembler language routine in SSS were designed to take advantage of the PDP10 processing features. It would not, therefore, be cost beneficial to convert the SSS to another machine. Update program modules would only have to be modified to create Structure, Nomenclature, Index and Search files using data set names unique to the new data base. The retrieval program would have to be modified to be able to switch to an additional data base.

E. User Support

Users of SSS are provided with a reference manual that describes in detail the various search strategies that can be employed. A comprehensive set of examples showing the terminal input and output for typical search queries is also provided in the manual. However, the user must be quite knowledgable in chemical structures to obtain maximum benefit. One day training sessions are available to SSS subscribers for $50. To obtain technical assistance, the user is provided with the telephone number of a representative of the Company responsible for the maintenance of the system, Fein-Marquart Associates.

F. Future Plans

The most ambitious plans for SSS include making a greater number of data bases available for chemical substance searching. Fein-Marquart under contract to EPA is currently in the process of making the MERCK Index available. Other data bases include Cambridge X-ray Crystal file, CPSC Chemicals, AEROS SOTDAT file, Las Vegas Chemical Spills, EPA Pesticide Standards, STORET Data Base, EPA Pesticides (inactive), EPA Pollutants in Drinking Water, EROICA, Gas Phase Proton Affinities, NCI-SRI Industrial Chemicals, NCI PHS-149, NIMH Psychotropic Drugs, and WHO International Name file of Drugs.

Fein-Marquart is currently working under contract to EPA on a probject to combine the separate sets of files into one set of files for SSS, to include:

- structure/nomenclature file
- index to structure/nomenclature file
- search files

This will enable the user to specify any number of files to be included in the search query. As more data bases are integrated, this will reduce the amount of redundant storage for CAS registry numbers that are cited in more than one data base. Included in this modification is the addition of a locator to indicate the data base in which information about this substance can be found. This locator will be printed when the user issues a display command for a temporary file. The capability will also exist to do away with the complete generation of the files that is currently required for each update. Instead, only changes would be applied to the existing SSS files. This would significantly reduce update run times as the data base increases in size.

Fein-Marquart also plans to implement the capability of saving the contents of a particular temporary file for use in some future terminal session, and to implement a nomenclature search capability. Incorporating the alternate CAS Registry Number into SSS is also being considered.

3.2.3 CIDS - Chemical Information and Data System

A. Information Content

CIDS is a total computerized system for the input, storage, and retrieval of chemical compounds and associated information. The data base contains about 50,000 compounds of interest to the U.S. Army. The system is located at the Aberdeen Proving Grounds in Edgewood, Maryland.

A user can search the data base off-line or on-line to identify chemical substances using an extensive set of molecular formula and structural fragment keys. To use CIDS effectively, a user must possess a thorough knowledge of chemical structures.

Table 3-3 is a list of data elements contained in CIDS. Each compound is identified by name, molecular formula, structural keys, and a set of identification numbers. Associated with each compound is a series of subject matter codes that categorize the information content of articles referencing the compound, and codes for the publication containing the citations.

B. Organization and Structure

The CIDS system utilizes an inverted file structure for rapid response to a query. The data base consists of three files:

- Data file
- Locator file
- Index file

The Data file consists of unit records containing information for each chemical substance in CIDS. Each record consists of a header containing the compound's CIDS registry number; data descriptors; and data flags followed by the data blocks described by the header. The header and all of the blocks are variable in length. Physical records consist of unit records which are blocked to tracksize length and can span physical records to increase input-output speed.

TABLE 3-3

CIDS DATA ELEMENTS

NAME	SEARCHABLE	PRINTABLE
Internal ID Number	X	X
CAS Registry Number (when available)		X
Experimental Agent Number		X
Other Identification Numbers		X
Molecular Formula	X	X
Molecular Formula Fragments	X	
Two Dimensional Structure Representation		X
Chemical Names		X
Structural Keys (codes)	X	
Subject Matter Codes		X
Bibliographic Citations		X

The Index file consists of an entry for each unique molecular formula and structural fragment key that is used to provide information about a chemical substance during the retrieval process. The appropriate keys are assigned to each chemical substance by a computer program during the data base update phase. Associated with each unique key is a pointer to a list of addresses.

This list of addresses is contained in the Locator file. The list consists of disk location information for each chemical substance to which a particular key was assigned. The unique disk location information for each chemical substance provides a means of structuring temporary file results of intersected and/or merged queries requested by the user. It also provides a direct access to the data file that decreases input-output time when searching or displaying the chemical substance unit record.

CIDS is updated on an as-needed basis by an in-house staff at Edgewood Arsenal. Information about each chemical substance is entered via punched paper tape produced by a teletypewriter equipped with special chemical type fonts. This information consists of identification numbers, molecular formula, structural formula (diagramatically represented), nomenclature, document references, and non-structural keys describing the contents of document references.

The paper tape images are transcribed onto magnetic tape and analyzed. The structural diagram is converted to a connection table (unique to CIDS) and the structure and molecular formula are verified for consistency and chemical veracity. The connection table is transformed to a compact linear code for storage. The structural representation and associated text are also stored in a highly compact form. A registry system assigns CIDS identification numbers to new compounds and updates or corrects information in previously registered compounds.

The molecular formula and structural keys are automatically generated by a computer program from the molecular formula and structure diagram input from the teletypewriter. This process results in the Index and Locator files.

During each update, the three data base files are totally recreated. This procedure is adequate for the current size of the data base (about 50,000 compounds) but would pose a problem if the size of the data base increased dramatically.

C. Retrieval Process

There are currently five terminals that access CIDS. All of these terminals are located at Edgewood. The terminals are TTY 37's that have been modified to include special chemical symbol keys. These keys are used to display a more chemically precise structural representation of a chemical compound. The computer software only allows two users to access CIDS simultaneously.

The basic search and retrieval stragety for CIDS is based on an extensive set of molecular formula and structural fragment keys designed to selectively retrieve chemical substances requested by a user.

A typical search consists of two phases. In the first, the user specifies the appropriate molecular formula and/or structural fragment keys to be employed in the search and the restrictions to be imposed on these keys. In the optional second phase, those compounds matching the first phase requirements are accessed and tested further with molecular formula requirements and/or atom-by-atom structural search descriptors. The user has the option of printing on-line or off-line the information about each compound retrieved.

The molecular formula keys follow the Hill format, which contains the total atom content of a compound. The molecular formula may be expressed either qualitatively or quantitatively. The qualitative method merely indicates the presence of a particular atom, whereas the quantitative method specifies the exact number of atoms. CIDS does not employ keys for nitrogen (N) or oxygen (O), but does provide specific keys for deuterium and tritium. CIDS also provides a General Metal key that is used to specify that the formula contains at lease one atom of any metal.

The Structural Fragment keys consist of a set of about 700 predefined codes assigned to chemical substances by a computer program during the update phase. These codes, which are categorized by structural characteristics, include information about the following properties:

- Organic - Inorganic
- Acyclic - Cyclic
- Extracyclic
- Number of Cyclic Nuclei
- Direct Attachments
- Generic Cyclic Nuclei
- Specific Cyclic Nuclei
- Functional Group (specific, nonspecific, diatomic, nonatomic)
- Hydrogen Radical
- Metal Cation
- Inorganic Anion
- Abnormal Mass (isotope)

During phase one, the user specifies the desired keys and combines them in the desired way using logical connectives. The Index file and Locator file are searched and the list of addresses are intersected and/or merged to produce a list of addresses that satisfy the search criteria. The user then has the option of further refining the search by imposing restrictions on the molecular formula and/or conducting a substructure search. In either case, information is obtained by searching the data file for each compound retrieved in phase one.

The molecular formula statement provides a rapid technique for limiting the counts and types of elements of all the retrieved compounds. It enables the user to define a range of possible formulas that cannot be economically stated through the molecular formula keys.

The substructure search allows the user to search for structural features which are not specifically tagged by the structural fragment keys. It is accomplished by doing an atom-by-atom search of each compound retrieved to determine if the substructure fragment specified by the user does in fact exist in any of the compounds. Such searches require a great deal of processing time and should be avoided for large numbers of compounds.

The user can refine the query in these two ways until satisfied with the results. At that time the information about each retrieved compound can be printed on the terminal or off-line. There is no user option as far as printing format is concerned; all the information about a particular compound is displayed. Both on-line terminals and off-line printers require equipment with special chemical symbols for structure representation.

As an alternative to the above search strategy, the user has the capability of displaying specific compound information by specifying the CIDS internal identification number or the CAS registry number when it exists. Besides the on-line interactive search capability, the user can fill out preformatted query specification forms which are converted to on-line retrieval-like commands on punch cards and submitted in a batch program.

D. Hardware and Software

CIDS is processed using a Univac 1108 under EXEC 8. Nonstandard hardware required include a printer, typewriter, and terminals equipped with special chemical structure symbols.

Retrieval of CIDS information is accomplished using a software package developed by Fein-Marquart Associates. The software was designed to allow a great deal of flexibility in implementing extensive revisions. CIDS is programmed almost entirely in Fortran V for a Univac computer. This Fortran level provides for bit manipulation and user-controlled I/O overlap. The retrieval software includes an overlay structure to be run in 50K words of core storage.

CIDS was designed to allow for functional growth with minimal interference with continuing system operation during the revision period. Data base records are variable length and each record element is identified by a system of data descriptors. To add another data element to the record, a new data descriptor must be defined and the update program modified to process this new element.

E. User Support

CIDS is supported by a comprehensive set of manuals that describe in detail the method for generating a query and the meaning of each of the molecular formula and structural fragment keys. However, a thorough knowledge of chemical structures is required to understand portions of the manual. There are no formal training sessions available to educate a CIDS user.

F. Future Plans

Currently, there are no planned enhancements or modifications to the CIDS hardware or software. The support staff is trying to broaden the user base by identifying potential new users and disseminating information about CIDS.

3.2.4 DR&D CIS - Drug Research and Development Chemical Information System

A. Information Content

The DR&D CIS is a total input, storage, retrieval, and monitoring system that supports the Cancer Treatment Program of testing compounds for anti-cancer activity within the National Cancer Institute (NCI). DR&D consists of two subsystems. The first builds and maintains the files that correspond to the structural and non-structural data about each compound. The second subsystem provides an on-line search and retrieval capability for certain data elements in these files. The following discussion analyzes only those aspects pertinent to the chemical substructure search capability portion of DR&D.

The data base contains confidential information on cancer related drugs; therefore it is only accessable by a limited number of users within NCI.

Table 3-4 lists the data elements contained in DR&D and indicates whether the elements are searchable and/or printable.

TABLE 3-4
DR&D CIS DATA ELEMENTS

NAME	SEARCHABLE	PRINTABLE
Extended Molecular Formula		
Molecular Formula	X	X
Molecular Formula Fragments	X	
Number of Rings	X	
Chemical Keys		
Augmented Atom	X	
Ring Elemental Analysis	X	
Nucleus	X	
Ring Size	X	
Elements	X	
Structural Feature Codes	X	
Two Dimensional Structure Representation	X	X
Internal ID Numbers (NSC & TID)	X	X
CAS Registry Number (50%)		X
Chemical Names		X
Chemical Properties		X

B. Organization and Structure

The DR&D CIS data base consists of four files:

- Structural Data
- Nonstructural Data
- Extended Molecular Formula Index
- Key Index

The Structural Data file contains connection table information associated with a particular chemical substance and is accessed when conducting an atom-by-atom search or when displaying the two-dimensional structure diagram on the terminal. The nonstructural data file contains all of the other information associated with a chemical substance. The Extended Molecular Formula Index and Key Index files are inverted lists of NSC numbers (internal ID numbers used to reference a chemical substance) that are associated with a particular search key. These two files are accessed when conducting a full structure or substructure search.

DR&D is maintained by the Chemical Abstracts Service (CAS). CAS receives chemical information to be input to the system on a Compound Submission Form (CSF). Data on the form that contain no structure information is added to the Nonstructural Data file. The structure diagram on the CSF is entered into the CAS Registry System TIP program via chemical typewriter. The registry system interfaces with DR&D to determine whether a structure already exists in the system and to assign NSC numbers to new compounds.

The Graphical Data Structure (GDS) formatted structural diagram is used to construct a connection table for the Structural Data file. The CSF information is also used to construct the Extended Molecular Formula and Chemical Keys Index files.

C. Retrieval Process

Any teletype-compatible terminal can be used to conduct a query session against DR&D. The system will also support a graphic-type terminal which provides vector graphics to produce a clearer visual representation of a two dimensional chemical structure diagram. Currently, three Hazetine 2000's and a Textronic 4200 are used to access DR&D.

Three methods of querying DR&D are available to the user:

- Full Structure Search (FSS) using the Extended Molecular Formula (EMF)
- Substructure Search (SSS) using Chemical Keys
- Direct Retrieval by Internal Identification Number (NSC & TID)

The Extended Molecular Formula includes molecular fragments, number of rings, and the number of direct attachments to rings. The Chemical Keys include:

- Augmented Atom - Specifies characteristics of a central atom. Each key specifies the element type of the central atom, element type of each attached atom, and bond type of bonds to each attached atom.

- Ring - Contains the element types and counts of all atoms in the ring path, two unsaturation counts, sum of alternating and delocalized charge bonds, and the sum of double, triple and tantomer bonds.

- Nucleus - Same as Ring but for the entire cycle nucleus.

- Ring Size - Indicates size of every ring.

- Elements - Indicates code for each element plus deuterium and tritium.

- Miscellaneous Keys - Associated structural feature codes. These keys can be combined using logical connectives to create the desired search query.

Before querying the DR&D data base the user must specify the information which is to be displayed about each substance.

After the user has indicated the method of querying and the search criteria, the retrieval system searches the appropriate index file and produces a list of NSC numbers of compounds that match the query requirements.

The user then has the option of performing an interactive atom-by-atom search of all retrieved compounds using a predefined substructure. The substructure must be defined and saved prior to executing the search command. The substructure is a two-dimensional graphic representation that is developed on the terminal using set of interactive commands to specify rings, atoms, bonds, chains, links, etc. A display command is used to print the substructure on the terminal for user verification. If this option is specified, the retrieval program matches the substructure against the appropriate connection tables on the Structural Data file. The Structural Data file is accessed using the list of NSC numbers that satisfied the initial query; the result is a list of NSC numbers which is a subset of the initial query results.

At this point, the user can do any of the following:

- print results on terminal
- print results off-line
- browse through results
- iterate further on results (or retrieve different data)

D. Hardware and Software

DR&D is processed through a contract with Chemical Abstract Services in Columbus, Ohio, using an IBM 370/168 with VS. Copies of the data base exist on disk packs, microfilm and computer-generated CHEM cards.

On-line search and retrieval sessions for DR&D are processed using a software package developed by the University of Pennsylvania. The program is written in PL/1 and takes up about 1.2 million bytes of real core storage. It was designed to be operated in a virtual operating system so the amount of core actually used at any given time is considerably less.

One drawback of the software is that it is not re-entrant, which means that each user requires a copy of the search and retrieval program. This is not a disadvantage in the current environment due to the low number of people that actually access the system. This situation probably would have to be changed if the number of users were to increase greatly.

The search and retrieval software and user interface is provided through TCAM (telecommunication access method). Access to the computer is protected by system of user-supplied passwords.

E. User Support

DR&D provides the user with a reference manual that explains the various search strategies available. Since there are only a limited number of users, no formal training sessions exist. User expertise is gained through in-house training and trial-and-error.

F. Future Plans

Future plans for DR&D include:

- Two-dimensional structure display for the entire file with more flexible on-line output
- Enhancing the user's manual by describing more examples of search queries
- Developing a report generator to combine structure and nonstructure output.

3.3 General Requirements for a Chemical Structure/Nomenclature System

This section presents the system requirements for implementing a comprehensive Chemical Structure/Nomenclature System (CSNS). These specific requirements are based on the conceptual requirements presented in the Mitre report. These requirements are not intended to be all inclusive, but rather a guide to be used in implementing the actual Chemical Structure/Nomenclature System.

3.3.1 Subject Matter and Scope

The purpose of an on-line CSNS will be to provide:

- A central storage capability for identifying information on as many chemical substances as possible
- As much identifying information about a chemical substance as possible
- An efficient on-line search capability against the identifying information
- A locator to indicate which data bases contain information about a chemical substance

The CSNS will consist of those chemical substances contained in data bases referenced by the locator. At a minimum, these will include data bases identified as part of the CSIN. Other data bases relevant to the study or regulation of chemical substances will also participate in CSNS. Both computerized and manual information systems will be included. CSNS will act as a focal point for users in government, universities, and private industry to identify chemical substances in a wide variety of areas, e.g., drugs, pesticides, and industrial chemicals.

3.3.2 Data Elements

The data elements presented below comprise a comprehensive set of elements that are desirable for a CSNS. These data elements will provide the user the greatest amount of flexibility in identifying or analyzing a chemical substance. Table 3-5 is a list of the required data elements; a description of each data element follows:

TABLE 3-5

DATA ELEMENTS COMPRISING A CHEMICAL STRUCTURE/NOMENCLATURE SYSTEM

NAME	SEARCHABLE	PRINTABLE
CAS Registry Number	X	X
Alternate CAS Registry Number	X	X
Chemical Names	X	X
Name Fragments	X	
Molecular Formula	X	X
Molecular Formula Fragments	X	
Wiswesser Line Notation (WLN)	X	X
Two-Dimensional Structure Representation (connection table)	X	X
Number of Rings	X	X
Ring Size	X	X
Ring Size Fragments	X	
Ring Elemental Analysis	X	X
Ring Elemental Analysis Fragments	X	
Structural Feature Keys (screens, codes)	X	
Total Molecular Weight	X	X
Total Number of Atoms	X	X
Component Line Formula	X	
Component Line Formula Fragments	X	
Data Base Locator		X
Data Base Subject Category	X	X
Data Base Data Element	X	X

CAS REGISTRY NUMBER - A unique nine-digit number used to identify a chemical substance. These numbers are assigned by Chemical Abstracts Services. Each chemical substance in CSNS will have a CAS registry number.

ALTERNATE CAS REGISTRY NUMBER - Registry numbers are assigned using an algorithm based on chemical structure. Occasionally, registry numbers are assigned before complete structure information is available; therefore, when a structure is fully defined, another more preferred number is assigned. A cross reference capability will exist to relate former numbers to new numbers.

CHEMICAL NAMES - Chemical substances are identified by a wide variety of naming systems in various data bases. Both systematic and generic naming conventions are employed. The comprehensive CSNS will include all names used by the various data bases referenced in the system. These naming conventions will include:

. CAS Preferred Name (7th, 8th and 9th collective index)
. Trade Names
. Synonyms
. IUPAC Name
. Generic Names
. Trivial Names
. Other Names

NAME FRAGMENTS - This field will enable the user to search on all meaningful parts of a chemical name. For consistency, the CAS name fragmentation scheme will be used.

MOLECULAR FORMULA - This field indicates the complete molecular formula of a chemical substance. All elements except hydrogen are specified together with a count of each element.

MOLECULAR FORMULA FRAGMENTS - This data element separately indicates each element name and occurrence number in the molecular formula.

WISWESSER LINE NOTATION (WLN) - This field is a unique and unambiguous representation of a chemical structure in a single character string using a standardized arrangement of numerics, alphabetics, and special characters.

TWO-DIMENSIONAL STRUCTURE REPRESENTATION - CSNS will utilize the connection table supplied by CAS to represent a two-dimensional structure diagram of a chemical substance. Such a connection table will contain data items describing the graph, nodes and bonds of each element in the chemical substance. Connection tables will provide a substructure searching capability (i.e., atom-by-atom search) for a complete or imbedded structure that matches a user-defined query developed at a terminal. These tables will also permit a user to display the two-dimensional structural representation of a substance at a terminal.

NUMBER OF RINGS - Indicates the number of rings in a ring system.

RING SIZE - Indicates the size of all the rings in a ring system.

RING SIZE FRAGMENTS - Indicates the ring size of each part of a ring system.

RING ELEMENTAL ANALYSIS - Indicates all element names and the number of occurrences of those elements that are contained in all the rings of a ring system.

RING ELEMENTAL ANALYSIS FRAGMENTS - For each ring of a ring system, this data item indicates all the element names and the number of occurrences of those elements.

STRUCTURAL FEATURE KEYS - CSNS will utilize the CIDS codes developed by the Army at Edgewood Arsenal. These predefined codes (about 700) are used to identify a comprehensive set of chemical characteristics. This data element will identify which of these codes are applicable to the particular chemical substance.

TOTAL MOLECULAR WEIGHT - Indicates the molecular weight, excluding hydrogen atoms.

TOTAL ATOM COUNT - Indicates the total number of atoms, excluding hydrogen.

COMPONENT LINE FORMULA - Indicates the atom sequence of all rings in a ring system.

COMPONENT LINE FORMULA FRAGMENTS - Indicates the atom sequence for each ring in a ring system.

DATA BASE LOCATOR - Indicates which data bases contain additional information about the chemical substance.

DATA BASE SUBJECT CATEGORY, DATA BASE DATA ELEMENT - Associated with each data base locator will be the subject category and data element name found in the Chemical Data Bases Directory (Section 2.2.3). This will enable the user to limit a search query to only those data bases that contain information about a certain subject category and/or data element. In order to minimize storage requirements the subject category and data element will not be associated with each chemical substance, but with each data base locator. A further description of these data items can be found in the Search and Retrieval Capabilities, and File Organization and Structure sections.

3.3.3 Search and Retrieval Capabilities

CSNS will provide the user with a broad range of search and retrieval capabilities in order to satisfy the wide variety of queries that will be inherent in such a comprehensive system. The following data elements discussed in the previous section will be searchable by the user:

- CAS Registry Number
- Alternate CAS Registry Number
- Chemical Names
- Name Fragments
- Molecular Formula
- Molecular Formula Fragments
- Wiswesser Line Notation
- Two-Dimensional Structure Representation
- Number of Rings
- Ring Size
- Ring Size Fragments
- Ring Elemental Analysis
- Ring Elemental Analysis Fragments
- Structural Feature Keys (screens)
- Total Molecular Weight
- Total Number of Atoms
- Component Line Formula
- Component Line Formula Fragments
- Data Base Subject Category
- Data Base Data Element

When searching the Molecular Formula Fragments, the user will be able to specify the presence of a particular chemical element in a compound, the exact number of occurrences of an element, or a range of occurrences numbers of an element. Similarly, when searching the total molecular weight or total number of atoms, the user will be able to specify an exact number or a range of numbers.

The user will also be able to construct or "draw" a two-dimensional representation of a query structure at a terminal using a series of interactive commands to specify rings, bonds, chains, etc. This query structure can then be used to search the data base for an embedded or exact structural match.

The computer software utilized for searching and retrieving will accommodate both the novice and experienced user. Since it is anticipated that CSNS will be accessed by many novice users, the interactive terminal conversation will prompt the user through the search and retrieval process. On the other hand, experienced users will be able to construct complex shorthand commands with little or no prompting. This will enable the experienced user to bypass the time-consuming user prompting.

To aid an inexperienced user, CSNS will provide a series of "help" commands that will explain how the system operates, describe the various data elements in the data base, and describe the structure of the various commands used to access the system.

During a typical search and retrieval session, the user will first specify a set of search criteria. The search criteria will be matched against chemical substances in the data base. The results of a query will be stored and "saved" as a list of chemical substances that satisfied the search query. The number of matches will be returned to the user. To accommodate multiple searches within the same session, a number of "save" files will be made available. Each save file will have a unique identifier so that it may be referenced in a later query.

Once the results of a search are returned to the user, he/she may:

- refine the search query
- display the results of the search
- initiate a new query

A search query can be refined by combining the contents of various "save" files using logical connectives (AND, OR, NOT). This will result in the creation of another "save" file.

The user will be able to display all the data elements of each chemical substance in a "save" file, or optionally specify that only certain data elements be printed.

A special query refinement will provide the user with the option of limiting the data base locator to only those data bases that contain information about a specified subject category and/or data element. For instance, this procedure would enable a user to determine what data bases contain information about a certain class of chemicals and also contain information about transportation control and fate in the environment. The class of chemicals would be identified using the procedures previously described and the results stored in a "save" file. The special query refinement will enable the user to limit the data base locator of each substance to those that contain information about transportation, control and fate in the environment. For each chemical substance in the "save" file, the search and retrieval program obtain the data base locator and search a separate file containing subject categories and data elements for each data base locator. If a matching subject category and/or data element is found, the data base locator will be saved. This procedure will eliminate the need for the user to make a separate access to the Chemical Data Bases Directory. In Section 3.5 Sigma Data will explore the possibility of a direct link to the Chemical Data Bases Directory to perform this function.

3.3.4 User Access

Access to CSNS will be through on-line conversational terminal sessions. Because of the anticipated wide variety of users, any TTY-compatible terminal will be supported. There will be no requirements for special terminals using vector graphics, or special chemical symbols; however, optional modules may be included to support such terminals. CSNS will be made available to all potential users. Users may exist, in the Federal government, state government, universities, private institutions, and private industry.

3.3.5 User Support

A complicated system such as CSNS requires a well written user's reference manual to aid the user in understanding the procedural and operational aspects of the system. The manual will include a clear description of the data base contents, the commands available to the user, and the way in which the commands are used to construct search queries and display results. Examples of actual terminal sessions that illustrate typical searches and retrievals will also be included in the manual.

Since many of the users of CSNS will not possess a thorough knowledge of chemical structure and searching capabilities, formal training sessions will be developed and conducted on an as-needed basis. These training sessions will introduce the user to the CSNS data base and explain the intricacies of constructing a query and the processes involved in constructing a search and retrieval. Actual terminal usage by class members will greatly reinforce their understanding of the system.

A user support staff will be available to answer any technical or procedural questions that may arise. Users having difficulty will be able to telephone the support staff to receive advice or answers to a particular problem.

3.3.6 Maintenance and Update

In order to maintain consistency with participating data bases, CSNS will be updated according to the update frequency of each of the participating data bases. During an update of a participating data base, substance identification information will be extracted that reflects all changes since the prior update. This extracted information will be sent to Chemical Abstract Services and matched against their data bases. A standard set of substance identification information about each chemical is extracted from the Chemical Abstract Services data bases and sent to CSNS where it is transformed into CSNS data base transactions. These transactions will then be used to update the CSNS data base.

This procedure will enable users to identify chemical substances in a participating data base using a comprehensive set of search criteria even though the data base is only searchable by a limited number of search items. For instance, if a data base is only searchable by chemical trade name, then by matching the trade name with the Chemical Abstract Services data bases, other data elements such as CAS registry number, CAS name, molecular formula, connection table, etc. are associated with the trade name. The comprehensive set of CAS information is incorporated into CSNS enabling the user to query using CAS registry number, and molecular formula. Once the desired substance is identified, the trade name will be displayed and used to access the participating data base.

If the actual implementation of CSNS does not provide a direct link to the Chemical Data Bases Directory, the subject category and data element information for each data base locator will have to be extracted from the Directory and incorporated into CSNS.

3.3.7 File Organization and Structure

Since immediate responses to user queries will be emphasized in CSNS, the file organization and structure techniques employed will accommodate rapid search and retrieval of data. Machine dependent organizational characteristics that increase processing speeds will be employed.

Each record of the CSNS data base will contain information for a particular chemical substance. A structural capability will be provided to search CSNS using these searchable data elements specified in Section 3.3.3. Any type of indexing, table-lookup or inverted-type file structure technique will be acceptable. Since CSNS will have a moderate to heavy update requirement, the organization and structure will employ efficient techniques for changing existing records and adding new records, i.e., the data base should not have to be completely regenerated during each update.

If the actual implementation of CSNS does not provide a direct link to the Chemical Data Bases Directory, a separate search file will have to be developed that contains the subject category and data element information for each data base locator.

The use of a commercially available data base management system will be considered, but the following reasons would discourage its use:

- The file organization is not complex.
- The file organization, once defined, will be static.
- Search and retrieval capabilities will be predefined and limited, not generalized Ad Hoc browsing.

3.3.8 Hardware

The critical component in CSNS is the search and retrieval capability. In order to process the user query requirements discussed in Section 3.3.3, a large, sophisticated computer program will be required. This type of real-time interactive program requires upwards of 300,000 bytes of core storage to operate efficiently. A fast computer processing speed will be required to accommodate the large number of simultaneous users accessing CSNS. Input-output speeds must be fast to process the volume of data required for search and retrieval. Since CSNS will contain information on nearly 500,000 substances, and since most of the information is searchable (which adds a great deal overhead for search data) a large amount of disk storage must be available. Wide band communications lines with fast transmission speeds will be provided to accommodate the number of users accessing CSNS.

CSNS will support the least sophisticated terminal user. Therefore, a TTY-compatible terminal will be used to access CSNS.

3.3.9 Software

Any software, such as extracts and updates, which are developed for CSNS should be written in a high-level language. Languages such as PL/1, Fortran, and COBOL provide easier maintenance and faster development time. Search and retrieval software should be re-entrant in order to process many user requests simultaneously. All software will be written in modular form to facilitate the addition of new functions or changes to existing functions. Input-output functions will be separated in order to isolate any machine dependent code.

To increase access speed, input-output modules may be written in assembly language. Although this language is machine dependent, it takes advantage of particular hardware characteristics and provides a faster response time for search and retrieval.

3.4 A Comparison of Existing Capabilities to Generalized Requirements

This section compares the four existing systems discussed in Section 3.2 to the generalized requirements discussed in Section 3.3. It illustrates the ability of the existing systems to meet the set of requirements for a comprehensive chemical structure and nomenclature system. Table 3-6 shows the results of the comparison.

Significant features and capabilities associated with the generalized requirements were identified and a numerical score was assigned to each system. A value of five indicates that the system totally supported the feature or capability. A value of zero indicates that the feature or capability was not supported at all. Numbers of one to four indicate varying degrees of support. The scoring for data elements was determined by the amount of coverage in the data base and the ease with which it is searchable and/or printable.

The total score for each system at the end of Table 3-6 indicates that two systems, CHEMLINE and SSS/CIS, support most of the features of the Chemical Structure/Nomenclature System. The other systems, CIDS and DR&D CIS, fall short in supporting CSNS features. The close score between CHEMLINE and SSS/CIS indicates that the overall level of support is about the same. A closer examination of individual requirements indicates that the two systems provide varying degrees of support. The highest possible score for supporting all requirements would be 205. If the highest score for each requirement between CHEMLINE and SSS/CIS was added, the total would be 187. This suggests that if the features of CHEMLINE and SSS/CIS were combined, practically all of the CSNS requirements would be satisfied. Section 3.5 will discuss alternatives to combining the two systems or separately enhancing one of the two systems.

TABLE 3-6

A COMPARISON OF EXISTING CHEMICAL STRUCTURE AND NOMENCLATURE SYSTEMS TO GENERALIZED REQUIREMENTS

FEATURE OR CAPABILITY	CHEMLINE	SSS CIS	CIDS	DR&D CIS
	(5=highest	score	0=lowest	score)
CAS Registry Number	5	5	1	1
Alternate CAS Registry Number	5	0	0	0
Chemical Names	5	3	2	2
Name Fragments	4	0	0	0
Molecular Formula	5	5	5	5
Molecular Formula Fragments	5	5	5	5
Wiswesser Line Notation (WLN)	2	0	0	0
Two-Dimensional Structural Representation (connection table)	0	5	4	3
Number of Rings	5	5	0	0
Ring Size	5	5	0	0
Ring Size Fragments	5	5	0	0
Ring Elemental Analysis	5	5	0	0
Ring Elemental Analysis Fragments	5	5	0	0
Structural Feature Keys (screens, codes)	0	5	5	5
Total Molecular Weight	0	5	0	0
Total Number of Atoms	0	5	0	0
Component Line Formula	5	2	0	0
Component Line Formula Fragments	5	2	0	0
Data Base Locator	4	2	0	0
Range of Chemicals Covered	3	4	2	3
Number of Unique Chemicals	3	4	2	4
No Chemical Expertise Required to Access System	4	2	2	2
Easy to Learn Search Strategies	4	3	2	2
Draw a Structure Diagram at the Terminal	0	5	0	0
Substructure Searching (atom-by-atom)	0	5	5	5

TABLE 3-6
(Cont'd.)

FEATURE OR CAPABILITY	CHEMLINE	SSS CIS	CIDS	DR&D CIS
	(5=highest	score	0=lowest	score)
Extensive User Prompting	2	4	1	2
Search Results Can be Saved	5	5	2	0
Search Results Combined With Previous Searches	5	5	0	0
Search Results Refined by User	5	5	3	2
Rapid Response Time	4	4	2	2
System Available to Many Users	3	5	1	1
Simultaneous Access by Many Users	5	5	1	0
System Available 24 Hours a Day	3	5	3	3
No Terminal Constraints	5	5	0	5
Optional Terminal Capabilities	0	4	0	4
Comprehensive User's Manual Provided	5	4	3	1
Formal User Training Provided	5	0	0	1
User Support Staff Provided	4	2	1	1
Data Base Updated Frequently	3	3	2	4
Machine Independent Software	3	3	1	2
Software Easily Modified	4	3	2	1
TOTAL	149	154	57	66

3.5 Implementation of a Chemical Structure/Nomenclature System

Based on the general requirements discussed in Section 3.3, the existing capabilities discussed in Section 3.2, and the comparison of existing systems discussed in Section 3.4, Sigma Data developed three alternative plans for implementing a Chemical Structure/Nomenclature System (CSNS). An analysis of the comparison of existing systems to the general requirements of a CSNS indicated that two data bases, CHEMLINE and SSS/CIS, supported most of the desired capabilities. It was therefore decided to utilize these two systems in developing the three implementation plans. The three alternatives involve modifying one or both of the systems and include:

- Enhancing both CHEMLINE and SSS/CIS to support the capabilities of the CSNS
- Selecting a single system to be enhanced to support the capabilities of the CSNS
- Distributing the search capabilities of CSNS between CHEMLINE and SSS/CIS.

The following subsections include a generalized system information flow description of the CSNS, a detailed discussion of each implementation plan, and a summary of the advantages and disadvantages of each plan.

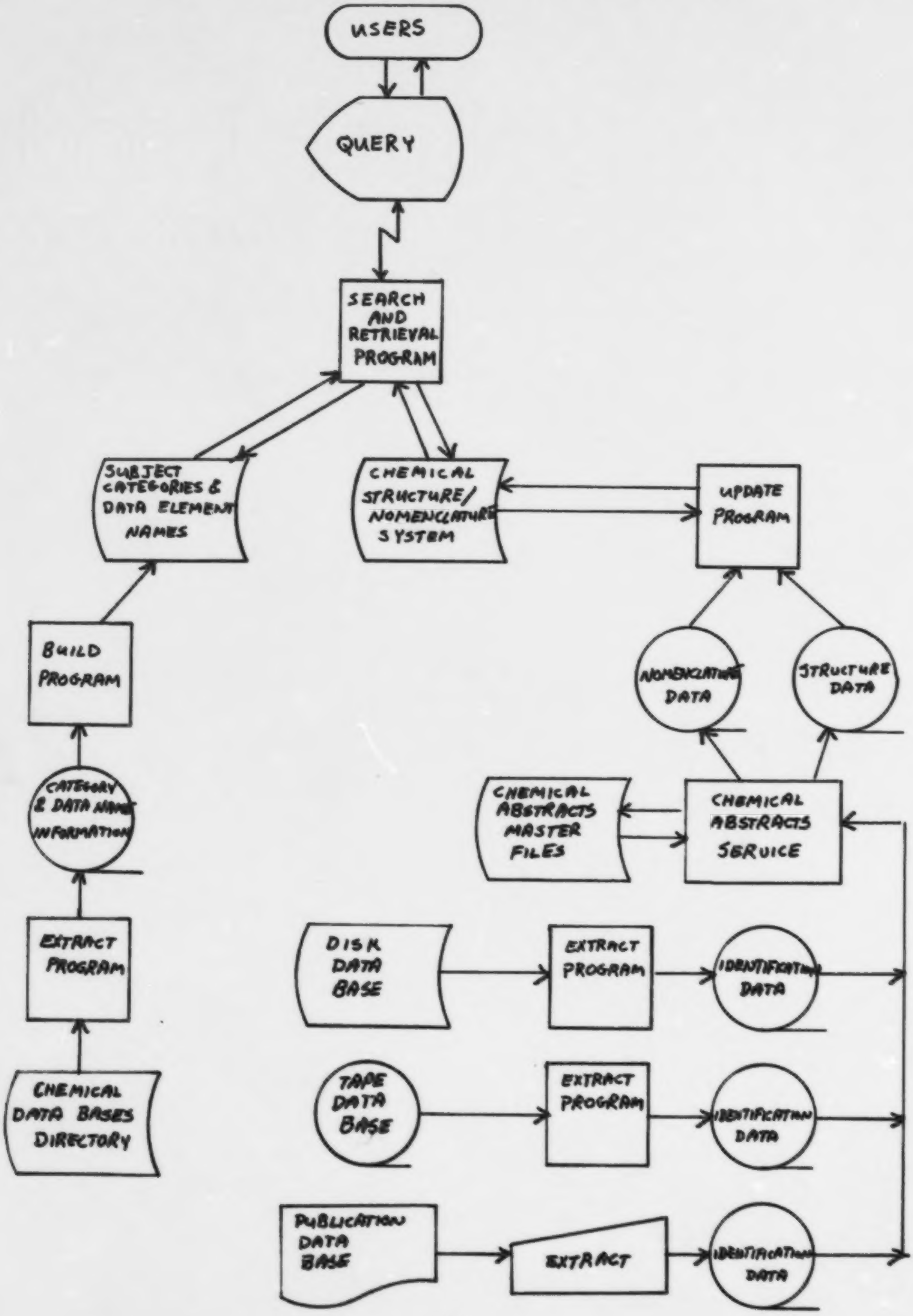
3.5.1 General System Description

Figure 3-1 illustrates a generalized system flow diagram of the proposed Chemical Structure/Nomenclature System that is not dependent on the implementation plan adopted. Data bases that have been selected for inclusion in CSNS would have identifying structure and nomenclature information extracted for each chemical substance and put on to magnetic tape. Once a data base is included in CSNS, periodic updates of the CSNS would only require information about chemical substances that have been added, changed, or deleted since the previous extraction.

The magnetic tapes containing structure and nomenclature data would be sent to Chemical Abstracts Service (CAS) where they would be matched against the Chemical Abstracts Master files. CAS would produce a set of records containing nomenclature and structure information in a standard format for each chemical substance matched. Therefore,

FIGURE 3-1

CHEMICAL STRUCTURE/NOMENCLATURE SYSTEM FLOW DIAGRAM


USERS
QUERY
SEARCH AND RETRIEVAL PROGRAM
SUBJECT CATEGORIES & DATA ELEMENT NAMES
CHEMICAL STRUCTURE/ NOMENCLATURE SYSTEM
UPDATE PROGRAM
BUILD PROGRAM
NOMENCLATURE DATA
STRUCTURE DATA
CATEGORY & DATA NAME INFORMATION
CHEMICAL ABSTRACTS MASTER FILES
CHEMICAL ABSTRACTS SERVICE
EXTRACT PROGRAM
DISK DATA BASE
EXTRACT PROGRAM
IDENTIFICATION DATA
CHEMICAL DATA BASES DIRECTORY
TAPE DATA BASE
EXTRACT PROGRAM
IDENTIFICATION DATA
PUBLICATION DATA BASE
EXTRACT
IDENTIFICATION DATA

if a particular data base only identified chemical substances by a name, CAS would match the name and provide the necessary structure information associated with the chemical substance. The structure and nomenclature data would be on separate tapes and provide the update mechanism for CSNS.

The CSNS update program would process the tapes supplied by CAS, construct additional CSNS dependent data elements using the CAS supplied information, and update the data and indexing files that comprise the CSNS data base. The update procedure would also include maintaining a searchable file of subject categories and data element names for each data base included in the CSNS. This information would be obtained from the Chemical Data Bases Directory (CDBD) on a periodic basis. An extract program would retrieve the information from the CDBD and a build program would construct the appropriate search files for the CSNS.

The user would query CSNS through a question-and-answer dialogue with a search and retrieval program. Chemical substances could be identified using any of the search capabilities discussed in Section 3.3.3. Once a chemical substance or group of chemical substances have been identified, the user may request a list of data bases that contain other information pertaining to these chemical substances. The user may optionally request a list of only those data bases that contain information about certain subject category and/or that contain data relating to a certain data element. In this case the data base locators retrieved from the CSNS would be screened and refined using the subject category and data element name search file extracted from the Chemical Data Bases Directory. Once the desired data bases have been identified, the user would have to conduct separate search and retrievals to obtain the additional information desired.

3.5.2 Plan I - Enhance Both CHEMLINE and SSS/CIS

The first alternative for implementing a Chemical Structure/Nomenclature System (CSNS) involves enhancing the two systems, CHEMLINE and SSS/CIS, that support most of the generalized requirements discussed in Section 3.3 This particular approach was suggested in the Mitre report on user needs for a Chemical Substances Information Network (CSIN). Implementation of this alternative would result in two search and retrieval systems both of which would satisfy the requirements of the CSNS.

Since the two systems employ different search methodologies, they could be analyzed in a working environment to determine which one would provide the best services to the CSIN user. At the end of the analysis period, a single system could be selected and designated as the one to be used as the Chemical Structure/Nomenclature System.

Table 3-7 lists the specific tasks required to enhance both CHEMLINE and SSS/CIS and a time phased schedule for implementation. Tasks related to CHEMLINE include:

Task 1 - Print Structural Diagram

A capability would be developed to print a two dimensional representation of a chemical substance at a terminal. For this task, it is recommended that NLM implement its current plan in which it is considering using CAS supplied information to provide this capability to CHEMLINE users.

Under this plan, the connection tables, which are tabular representations of chemical structures, would be obtained from CAS for each substance on CHEMLINE and inserted into the Header File of the data base. A print module supplied by CAS would be incorporated into the ELHILL software to transform the connection table into a two dimensional structural representation that could be printed at the terminal.

Task 2 - Draw Structural Diagram

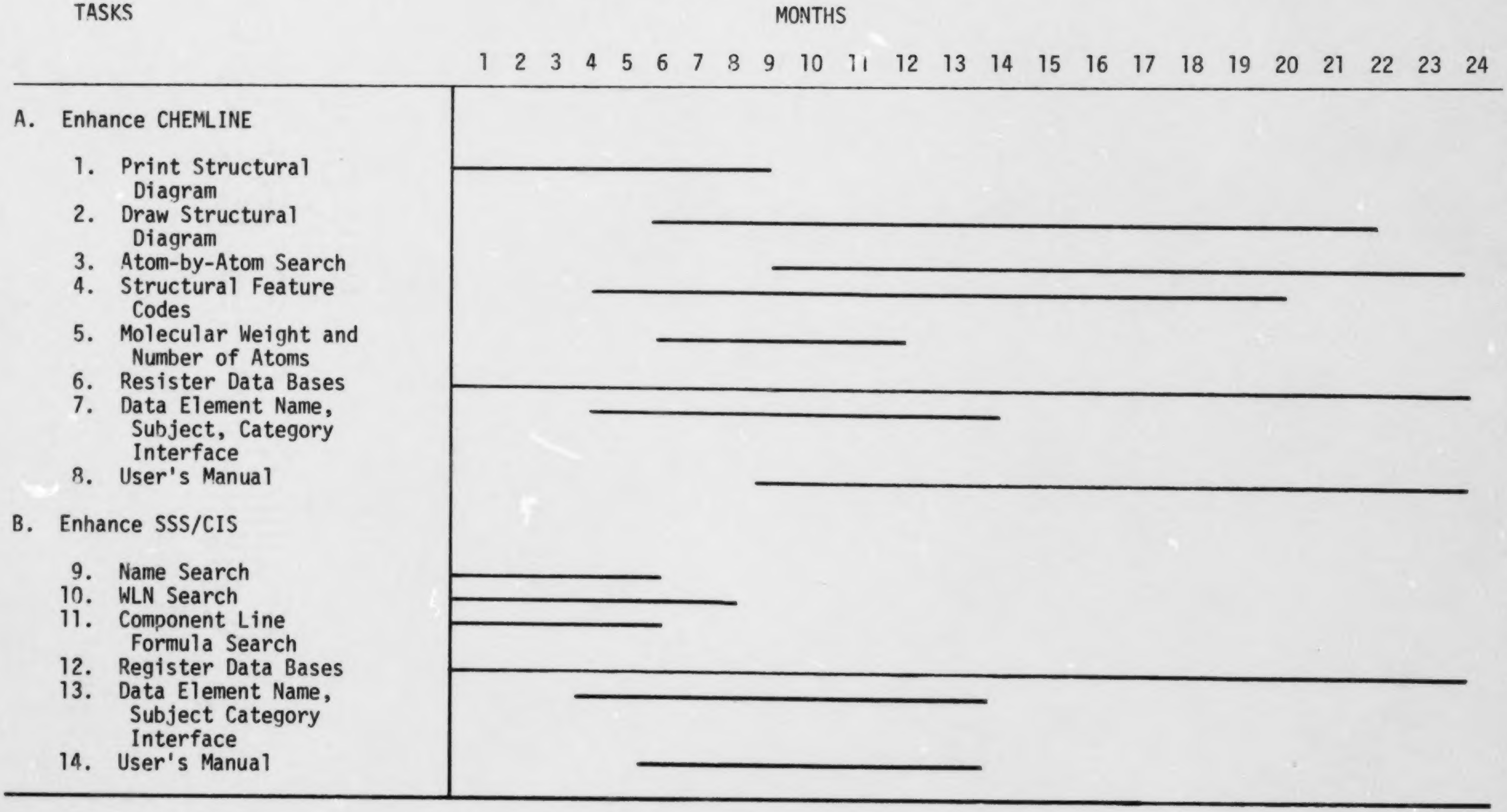
A user-promoted interface would be developed to allow a user to "draw" a two dimensional structural representation of a chemical substance. The diagram would possess the same characteristics (i.e. connection table) as the printed diagram discussed in Task 1. The user would be able to specify rings, chains, bonds, nuclei, elements, etc. through an interactive dialogue with the interface program. The user would also be able to display, save, or modify the diagram any time during the terminal session. The ELHILL software would be modified to provide a "draw" command and to provide a linkage between the "draw" interface program.

Task 3 - Atom-By-Atom Search

A search module would be developed to use the structure diagram developed in Task 2 to search a subset of the CHEMLINE data base. The connection table for each substance on the Header file would be compared to the connection table of the substance that was "drawn" by the user. The comparison would be for a complete structural match or a structural fragment match. The ELHILL

TABLE 3-7

CSNS PLAN I IMPLEMENTATION TASKS



software would be modified to provide an "atom-by-atom" search command and a linkage between the search module.

Task 4 - Structural Features Codes

A system of Structural feature codes similar to the CIDS keys used in SSS/CIS would be incorporated into CHEMLINE. The codes would be derived from the connection table of a chemical substance during the update process. The derived codes would be entered into the Header File and the appropriate Index and Posting Files entries would be generated for each key. The ELHILL software would be modified to search for and print the set of structural feature codes.

Task 5 - Molecular Weight and Numbers of Atoms

The CHEMLINE data base would be enhanced to include the molecular weight and total number of atoms data elements. These data elements would be determined from the molecular formula of a chemical substance during the update process. The data would be entered into the Header File and the appropriate entries would be made in the Index and Posting Files. The ELHILL software would be modified to search for and print these two data elements.

Task 6 - Register Data Bases

Data bases that are to be referenced by the CSNS must have identification information extracted and matched by Chemical Abstracts Service. As a result of the matching process, Chemical Abstracts Service would produce a nomenclature and structure file of required information in a standard format that would be used to update the CHEMLINE data base.

Initially, the approximately 45 component data bases of the Chemical Substances Information Network identified in the Mitre report should be registered with CHEMLINE. Since data base identification information extraction and CAS matching involves a great deal of effort, cost and time, a priority system should be developed to register the remaining data bases referenced in the Chemical Data Bases Directory.

EPA, in conjunction with SSS/CIS, will have matched nearly 40 data bases through Chemical Abstracts Service by the end of October 1977. Of these data bases, 17 are included in the component files of CSIN. The structure and nomenclature files produced for EPA could be used to register these same data bases in CHEMLINE, thus reducing a duplication of effort. Registration of the remaining CSIN component data bases should be a coordinated effort between EPA and NLM.

Task 7 - Data Element Name and Subject Category Interface

A special query refinement discussed in the CSNS Search and Retrieval Capabilities (Section 3.3.3) would enable a user to limit the results of a search to only those data bases that contain information about a certain subject category and/or data element. Implementation of this capability depends where the Chemical Data Bases Directory resides. If the CDBD is incorporated into MEDLARS, a program module could be incorporated into the ELHILL software that would access the CDBD directly for subject category and data element information using the data base locator information for appropriate chemical substances in the CHEMLINE data base.

If the CDBD is not incorporated into MEDLARS, the subject category and data element information would be extracted from the CDBD and entered into an indexed file that would be accessed by the CHEMLINE component of MEDLARS. A program module would be incorporated into the ELHILL software that would access the subject category and data element file to refine the data base locators for chemical substances in CHEMLINE. The ELHILL software would be modified to provide a new search command and the proper linkage to the search module. Implementation of this task would have to be coordinated with the implementation of the CDBD discussed in Section 2.4.

Task 8 - User's Manual

The CHEMLINE Portion of the MEDLARS User's Manual would be updated to reflect the additional enhancements. Training sessions would also be modified to include the new information available through CHEMLINE.

Tasks related to SSS/CIS include:

Task 9 - Name Search

In order to incorporate a chemical name search capability into SSS/CIS, appropriate Search files containing names and name fragments, as well as user-prompted search and retrieval dialogue would be developed. The search files would be constructed from the nomenclature file supplied by CAS during the SSS/CIS updating process.

Task 10 - WLN Search

In order to incorporate a WLN Search capability into SSS/CIS, a computerized algorithm would be developed to calculate the WLN for substances from the data on the Structure file supplied by CAS. The appropriate Search Files would be constructed as well as the user prompted dialogue necessary for search and retrieval.

Task 11 - Component Line Formula Search

Developing a component line formula search ability would require the construction of appropriate Search Files and software enhancements to provide a user-prompter query capability to retrieve the desired results. The component line formula would be determined from the connection table supplied by CAS for each chemical substance and applied to the Search File during the update process of SSS/CIS.

Task 12 - Register Data Bases

SSS/CIS consists of identification information for chemical substances and indicates other data bases that contain additional information about the chemical substance. In order for other data bases to be referenced by SSS/CIS, identification information must be extracted and matched against the Chemical Abstracts Service Master Files. CAS then produces a standard set of structure and nomenclature information that is used to update the SSS/CIS data base.

EPA will have matched nearly 40 data bases through CAS by the end of October 1977. Of these data bases, 17 are included in the component files of the Chemical Substances Information Network identified in the Mitre report. In order to provide a more complete information retrieval capability for the Office of Toxic Substances, an effort should be made to register the remaining component files of CSIN. Since the registration process is quite costly, a priority scheme should be developed for including other data bases in SSS/CIS. The registration process should be coordinated with NLM so that the CAS information could also be used to update the CHEMLINE data base

Task 13 - Data Element Name and Subject Category Interface

A special query refinement discussed in the CSNS Search and Retrieval Capabilities (Section 3.3.3) would enable a user to limit the results of a search to only those data bases that contain information about a certain subject category and/or data element name information would be extracted from the Chemical Data Bases Directory and used to construct the necessary Search files in the SSS/CIS data base. The SSS/CIS search and retrieval software would be enhanced to provide for user specifications to refine the search results and procedures for accessing the Search files. Implementation of this task would have to be coordinated with the implementation of the CDBD discussed in Section 2.4.

Task 14 - User's Manual

An explanation of the additional search and retrieval capabilities would be included in the current SSS/CIS User's Manual. Training sessions would also be enhanced to reflect these additional capabilities.

Under this plan, CSNS capabilities could be operational on SSS/CIS within 14 months, and operational on CHEMLINE within 24 months. The registering of data bases would be an ongoing operation throughout the implementation period, with a heavy emphasis on registering CSIN component files within the first 14 months. It is estimated that 65 man months of effort would be needed to enhance CHEMLINE and 16 man months would be needed to enhance SSS/CIS. This time would be divided between project managers, systems analysts and programmers. The greater level of effort required to enhance CHEMLINE is mainly due to the design and implementation of Tasks 2, 3, and 4, which include "drawing" a structural diagram, providing an atom-by-atom search, and providing a structural feature code search.

The cost to implement Plan I is estimated at $260,000 for CHEMLINE and $65,000 for SSS/CIS. This cost does not include computer operation time or the cost to register the CSIN component files. Since 17 out of the 45 CSIN component files have been registered by EPA already, the cost to register the remaining 28 data bases is estimated at $280,000.

3.5.3 Plan II - Select a Single System to Develop

The second alternative for implementing a Chemical Structure/Nomenclature System involves evaluating the two systems, CHEMLINE and SSS/CIS, that most closely meet the long-term requirements of a CSNS. The results of the analysis would be used to determine which of the two systems will best meet the needs of the Chemical Substances Information Network (CSIN) and should be enhanced to satisfy the general requirements of the CSNS discussed in Section 3.3. This approach would eliminate the duplication of effort involved in the enhancement of both CHEMLINE and SSS/CIS discussed in Plan I.

The first task in this plan is to conduct a detailed analysis and evaluation of CHEMLINE and SSS/CIS in order to determine which system would most economically satisfy the long-term requirements of CSNS. Since the CSNS may be required to accommodate as many as 500,000 chemical substances, the analysis should consider the systems ability to store and access up to this number of chemical substances in an economical manner.

The search and retrieval methodologies should be analyzed to determine what impact the increased number of substances would have on the amount of time required to conduct a search and the additional storage required to hold the indexing information. The systems ability to provide rapid response time to an increased number of users should also be analyzed. The evaluation should also include a detailed user needs analysis to determine what data elements are required for chemical identification and searching, and what types of user search strategies could be best employed in identifying a chemical substance.

The results of the analysis would include a description of the tasks and their associated costs required to enhance reach system to meet the long-term CSNS requirements. The description would be similar to the Plan I presentation. A decision could then be made to select the most appropriate system to be utilized by CSNS.

It is estimated that the analysis would require 6 man months of effort and cost $30,000.

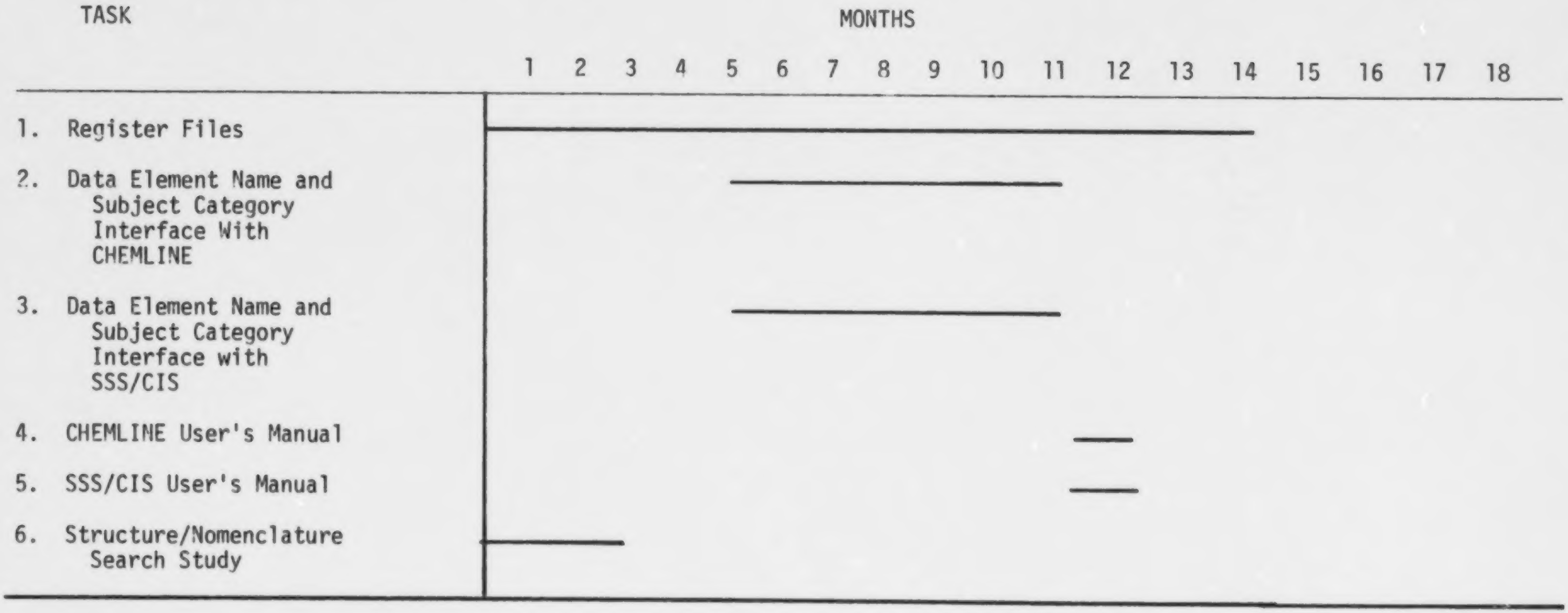
3.5.4 Plan III - Distribute Search Capabilities Between CHEMLINE and SSS/CIS

The third alternative for implementing a Chemical Structure/Nomenclature System involves the utilization of both CHEMLINE and SSS/CIS for different situations. Under this option, the various search strategies employed to identify chemical substances would be distributed between the two systems.

The two major means of identifying a chemical substance are through nomenclature searching and chemical structure searching. A comparison of CHEMLINE and SSS/CIS in Section 3.4 indicates that a well defined nomenclature searching capability exists within CHEMLINE, while a well defined structure searching capability exists within SSS/CIS. Under this option, each system would maintain its current search and retrieval capabilities. As future search and retrieval requirements are identified they would be assigned to one of the systems. A user would access one of the two systems depending upon the type of identification information

TABLE 3-8

CSNS PLAN III IMPLEMENTATION TASKS



known. For instance, an analyst at the Office of Toxic Substances that received a notification of risk about a certain chemical might access CHEMLINE using the chemical name to determine what data bases contained additional information about the chemical substance. A person conducting research concerning hazard assessment might access SSS/CIS using the structure of a particular substance to identify other chemicals that have similar structural configurations.

Table 3-8 lists the tasks required to implement Plan III and a time phased schedule for completion. A description of each task is given below:

Task 1 - Register Files

Data bases that are to be referenced by the CSNS must have identification information extracted and matched by Chemical Abstracts Service (CAS). CAS will then produce a nomenclature and structure file of required information in a standard format that will be used to update both CHEMLINE and SSS/CIS.

Initially, the approximately 45 component data bases of the Chemical Substances Information Network identified in the Mitre report should be registered to provide the Office of Toxic Substances the necessary data base locators. Since data base registration involved a great deal of effort, cost, and time, a priority system should be developed to register the remaining data bases referenced in the Chemical Data Bases Directory.

EPA in conjunction with SSS/CIS will have registered nearly 40 data bases through Chemical Abstracts Service by the end of October 1977. Of these data bases, 17 are CSIN components. The data bases registered by EPA could be used to update the CHEMLINE data base. Since both systems could be updated using the same CAS supplied information, registration of the remaining 28 CSIN component data based and any future data bases should be a coordinated effort between EPA and NLM.

Task 2 - Data Element Name and Subject Category Interface With CHEMLINE

This task is identical to Task 7 discussed for Plan I in Section 3.5.2.

Task 3 - Data Element Name and Subject Category Interface With SSS/CIS

This task is identical to Task 13 discussed for Plan I in Section 3.5.2.

Task 4 - CHEMLINE User's Manual

The MEDLARS User's Manual would be updated to reflect the additional CHEMLINE capability of refining a search according to subject category or data element name (Task 2).

Task 5 - SSS/CIS User's Manual

The SSS/CIS User's Manual would be updated to reflect the additional capability of refining a search according to subject category or data element name (Task 3).

Task 6 - Structure/Nomenclature Search Study

A study should be conducted to determine user needs for searching a chemical identification data base using structure and nomenclature strategies. The study should identify the particular data elements and their characteristics required for each strategy. The results of the study would be used to determine what searchable data elements should be added, deleted or modified in the CHEMLINE and SSS/CIS data bases to make them more responsive to user needs.

Under this plan, the only immediate enhancement required for CHEMLINE and SSS/CIS is the data element name and subject category interface with the Chemical Data Bases Directory and changing the user manuals to reflect the new capability. It is estimated that the level of effort required for each system would be 7 months and the cost would be $30,000. The structure/nomenclature study would take an estimated 3 months and cost $15,000. A total of 17 out of the 45 CSIN component data base have already been registered through Chemical Abstracts Service. The remaining 28 data bases could be registered within a 14 month period for a cost estimated at $280,000.

3.5.5 Advantages and Disadvantages of Each Plan

The three implementation alternatives are compared in order to identify their strengths and weaknesses. Table 3-9 is a summary of the advantages and disadvantages of each Plan.

Plan I involves enhancing both CHEMLINE and SSS/CIS to satisfy the requirements of the Chemical Structure/Nomenclature System. Both systems would provide the same information to the user, but would retain their individual search and retrieval methodologies. One advantage to this plan is that a user could choose the most convenient system and still receive the same results. Usage would be spread between two systems. This may become very important in the next few years when the number of anticipated users will increase substantially. User contention for search and retrieval

TABLE 3-9

A COMPARISON OF CSNS IMPLEMENTATION PLANS

PLAN	ADVANTAGES	DISADVANTAGES
I. Enhance Both CHEMLINE and SSS/CIS	· Test search methodologies in an actual operational environment · User can choose system to access · Usage spread between two systems · Less user contention, faster response time · Backup capability	· Duplication of effort in software development · Redundant storage and processing of the same information
II. Select a single system to be enhanced	· Eliminate redundant information in two systems · Eliminate duplication of effort in software development	· A single system must process all user requests · System selected must be enhanced with addition search strategies
III. Distribute search capabilities between CHEMLINE and SSS/CIS	· Usage distributed between two systems · Fine-tune existing search strategies · Software development kept to a minimum	· User limited to certain system based on type of search · User may have to maintain two accounts · Some redundant information carried on two systems · User may have to learn search and retrieval commands for two systems

services would decrease if two systems are available, thus decreasing the response time for user queries. If one of the systems is inoperative, the other system could act as a backup and absorb the extra workload, thus providing continuous service to the user. Under this plan, the different search and retrieval methodologies could be tested and evaluated under actual operational conditions. If one of the systems is unable to provide the required user services, it could be abandoned and the other system could take over full responsibilities.

One disadvantage of Plan I is the duplication of effort involved in software development. Each system must incorporate search and retrieval commands and processing that are already operational in the other system. Under this plan both systems must store the same information. Currently, this does not require a significant amount of storage because of the number of substances (around 200,000). However, in the near future, the number of substances may approach 500,000. Storage of identification information for each substance together with the storage requirements for the indexing information of the many data elements that can be searched in CSNS could be quite significant.

The second alternative involves selecting either CHEMLINE or SSS/CIS to represent the Chemical Structure/Nomenclature System in the overall Chemical Substances Information Network. Once the selection was made, the single system would be enhanced to meet all the user requirements for the CSNS. All data bases of interest to CSIN would be registered with the single system selected. This approach would eliminate redundant information being carried on two systems. It would also eliminate the duplication of effort required in Plan I to develop software for identical search and retrieval capabilities in two systems. However, the single system selected must incorporate software enhancements required to meet the search and retrieval capabilities discussed in Section 3.3.3.

The major disadvantage with Plan II is that a single system must process all users requests. It is anticipated that the number of users accessing CSNS will greatly increase in the next few years as well as the number of chemical substances contained in the data bases. As the number of users increase and the amount of indexing information that has to be searched to satisfy a query increase, the response time per query also increases.

The third alternative involves distributing the various searching capabilities for CSNS between CHEMLINE and SSS/CIS. CHEMLINE would specialize in search strategies using nomenclature while SSS/CIS would specialize in search strategies using chemical structure. System usage would have to select a particular system based on the type of search to be conducted. Each system would retain their current search capabilities and would not have to develop software to support search capabilities that already exist in the other system as Plan I requires. Each system could concentrate on fine-tuning their particular search capabilities in order to provide optimum support and response time to a user's request.

The major disadvantage with Plan III is that a user would not be able to conduct both a nomenclature and structure search during the same terminal session. A user would have to manually save the results of one search, switch over to the other system, and manually enter the results of the previous search in order to continue the search in the other system. There would be some redundant identification information about each substance in both systems, but duplicate indexing information for searching the various data elements would be eliminated. If the needs of a particular user required both nomenclature and structure searching, accounts with two systems would have to be maintained and the user would have to learn the search and retrieval command language for both CHEMLINE and SSS/CIS.

4.0 COMPUTER SYSTEM SECURITY

4.1 Introduction

The security of computer systems and their contents continue to be a topic of increasing concern. Security in computer systems involves the protection of the system from user, operator, administrative, and maintenance activities which could result in misuse of the system and its components. In the past, the major impetus for developing secure systems has come from the defense community. Currently, however, four factors have combined to motivate widespread interest in computer system security:

- Privacy
- Fraud
- Data Sharing
- File Contents Value

Privacy, the legal and social right of an individual to have personal data remain confidential, requires that personal confidential data be secure against unauthorized uses and users. The increasing evidence of computer-connected fraud and its huge costs have emphasized the need to restrict programmer access to systems resources. Data sharing has focused attention on computer system security as a major means of maintaining control of programs and data. Finally, the value of file contents highlights the need to protect stored data from unauthorized access and modification.

Recent trends toward the use of remote terminals, multi-programming, and concurrent use of system resources have increased the pressures to provide secure computer systems which protect against accidental or deliberate misuse of system resources.

The objectives of this section are to identify:

- Security threats to which systems such as the proprietary Toxic Substances Control Act Chemical Data System (TSCA CDS) are vulnerable

- Computer technology methods that can provide security against those threats

- Hardware and software requirements for implementing appropriate security methods

Section 5.2.4.3 of this report will use the information presented in this section to discuss the computer technology required to provide security for the TSCS CDS.

4.1.1 Background

The Mitre Corporation report Chemical Substances Information Network identifies security considerations for only one component system of CSIN, the Toxic Substances Control Act Chemical Data System.[1] Security is a major concern for the TSCA Chemical Data System because of the chemical industry trade secret data it will contain.

Exposure of trade secret data, such as new chemicals to be produced and the quantities of specific chemicals produced, could upset industry firms' competitive positions and lead to financial loss for some firms. In Congressional hearings on the Toxic Substances Control Act, industry representatives expressed great concern that these data remain confidential. The Mitre analysis reveals that much of the data reported by industry usually will be declared confidential.[2]

Because of the potential value of the data to industry competitors, and the potentially adverse effects of the Office of Toxic Substances regulations on the industry, it is not inconceivable that attempts will be made to compromise the integrity of the data base by destroying or modifying data, or by denying service to users at a crucial time. The TSCA CDS therefore must be secure in order to prevent unauthorized access to data and unauthorized manipulation of data.

4.1.2 Approach and Scope

This section of the report begins with a general discussion of security problems and solutions associated with computer technology. The discussion is based on a survey of published literature (1972 to present) related to computer system security. A list of references is presented at the end of Section 4.0.

---

[1] M. Bracken and J. Hushon, Chemical Substances Information Network. Volume II: Procedures for Handling Confidential Data Submitted Under the Toxic Substances Control Act. The Mitre Corporation, McLean, Va., June 1977.

[2] Ibid., p. 2-12.

In order to identify the computer technology security requirements for the TSCA CDS, it is first necessary to identify the computer technology security threats and safeguards applicable to most contemporary computer systems. These must then be reviewed to determine the methods of security which are applicable to the TSCA CDS, the way in which they can be implemented, and the software and hardware capabilities that are required by each method. Further, to ensure that computer technology security mechanisms are employed accurately and effectively, it is necessary to identify the security threats that are likely to occur in the TSCA CDS.

The discussion of the kinds of threats to which computer systems technology is vulnerable is divided into the major security threats and the users/intruders who typically present those threts. The basic ways in which computer technology security protects against threats, and the computer software and hardware utilized to implement the protection methods are then discussed.

Section 5.2.4.3 of this report will combine the generalized information on computer security problems and solutions with the specific security analysis and recommendations contained in the Mitre report. Analysis will result in identification of computer technology requirements for TSCA CDS security.

4.2 Characterization of Security Risks

4.2.1 Introduction

The first step in providing computer system security is to conduct a risk assessment, which involves identifying and prioritizing the threat of unauthorized access or degraded service. The likelihood of occurrence and the costs involved in each successfully executed violation determine the threat priority. Based on these data, accurate decisions then can be made concerning the concentration of the available effort, resources, and mechanisms to provide the best possible security.

Normally a risk assessment encompasses the entire computer installation. Volume II of the Mitre report thoroughly discusses the risks associated with physical, personnel and administrative security; this report specifically focuses on the security risks which are created or allowed by computer technology.

4.2.2 Types of Security Threats

There are two types of threats inherent in a computer system: accidental and deliberate. Accidental threats are those which are not motivated by malacious intent, but which usually result from user errors. For example, a user might enter an erroneous or inaccurately keyed password, or might accidentally request an unauthorized file or program. A user might also attempt to update a record to which he/she is permitted retrieval access only. Such user mistakes result in unauthorized exposure or manipulation of confidential data, or may cause the system to fail. Because these threats are accidental, they occur in random fashion and are not likely to be taken advantage of by the user. In addition, such problems are not severe because they can be solved by applying existing protection devices. Accidental errors are generally not a major security threat, although they should not be treated casually.

Other accidental threats are caused by system errors, which include incorrectly designed and/or implemented applications programs, systems programs, and operating systems. Although these problems generally involve reliability rather than security, security threats do occur, e.g., errors can expose confidential data to unauthorized access or manipulation, or can result in degraded service. Further examples of errors include computer programs, co-interference or interference with the operating system, temporary core storage in another user's space, and undocumented interrupts to communications lines. System errors pose significant security risks because they make it relatively easy to subvert the protection mechanisms.

Deliberate threats, both passive and active, are intentional attempts to compromise security by unauthorized observation, access, manipulation, or destruction of confidential data. A passive threat involves only hearing or observing data stream flow over communications lines; wiretapping is the principal example. Observers have no control over the data flow and rely on obtaining valuable information from raw data. The security risks posed by passive threats are associated primarily with confidential proprietary and personal data. The proposed TSCA CDS will be particularly susceptible to wiretapping because it will contain a great deal of proprietary industrial data which can be accessed from remote terminals. Exposure of trade secret data such as the manufacturing process or the quantity to be manufactured could severely upset a firm's relative competitive advantage, at a cost of millions of dollars in profits.

Active threats are more difficult to prevent because they involve utilizing the computer system's resources against itself. Deliberate intruders, including persons authorized to use the system in other ways, can use a number of techniques to obtain unauthorized access to the system. For example, a legitimate user from an on-line terminal could simply browse through unauthorized files from which he/she is not prohibited by protection mechanisms. In a system utilizing password protection, an unauthorized user might steal a security password and thus achieve access to confidential data. Some computer systems allow illegitimate user access by permitting an unauthorized user to connect his/her terminal to a communications port just after a legitimate user has disconnected but before the computer system has terminated the transaction. This type of illegitimate access may include not only access to the data but also to system resources to manipulate the data. Active threats therefore can be quite costly if successfully executed; however, they usually can be defeated by application of complex access control mechanisms.

Active threats which exploit the system programs and operating system are more difficult to counter. One type of deliberate threat utilizes system software and operating system idiosyncrasies and mistakes to trap data in unauthorized storage locations from which the data can later be retrieved. Similar software weaknesses can be exploited to penetrate the operating system and subvert all security mechanisms by gaining control of the operating system. Security then can be compromised by performing a core dump to expose the contents of the computer system. A user can also gain control of the operating system, usually in a supervisory or privileged state, to disable all protection mechanisms and add himself/herself as an authorized user, thus enabling future access to all data and system resources. At present, there is no known computer system that cannot be penetrated. Pentrations therefore present the most serious security risks and are the most difficult threats to combat.

4.2.3 Types of Users/Intruders Posing Risks

The groups of people who can be expected to initiate threats include:

- General System Users
- System Personnel
- Outsiders

The characterizations of these potential risks are based on the likelihood of group members attempting intrusion and the probability of successful execution.

General systems users (i.e., those at remote terminals) do not present major security concerns. Threats primarily result from accidental errors rather than from intent. Normally, users are limited in their knowledge of the system's design, operation, and weaknesses, particularly when users can access, retrieve, and manipulate shared data only via system applications programs. When data files can be updated or when user programs can be entered from a remote terminal, security risks increase substantially. In the first case, the accuracy and integrity of the data could be affected. In the second case, user programs could accidentally or intentionally access or manipulate unauthorized data, damage the data, or otherwise disrupt the system operations.

System personnel, both applications programmers and systems programmers, are potentially greater risks than users. because of their programming ability, knowledge of the system design and operation, access to systems manuals, and opportunities for system access. Recent evidence indicates that disgruntled former data processing personnel have been responsible for most of the costly fraud-related security violations. It is reasonable to assume that the situation will be similar for a system containing trade secret data.

Applications programmers present slightly less risk than do systems programmers, because they generally are limited in their knowledge of system operations details, and possess less sophisticated programming abilities. Applications programmers can, however, deliberately or accidentally write programs which result in unauthorized access, damage to the data, or disruption of service. An applications programmer responsible for updating a data base has the opportunity to make an extra copy of the file. Additionally, applications programmers are in a position to attempt to exploit the system software and/or penetrate the operating system.

"Outsiders" are those individuals who are not authorized to use a particular computer system. Outsiders can deliberately mount both active and passive threats to computer system data protection. In the proposed CSIN, the proprietary TSCA CDS is the most likely target of these threats primarily because it contains valuable trade secret information. Active threats such as guessing passwords and penetrating the operating system require a terminal which can access the computer system. Because proposed security regulations for the proprietary TSCA CDS include physical security for remote terminals, the security risk of active threats from an authorized terminal is low. The risk of an unauthorized terminal masquerading as a legitimate one is moderate.

The passive threat by outsiders is a much more serious risk, however. Wiretapping is relatively easy and inexpensive to employ against any system utilizing remote terminals and communications lines. Access to industry competitors' trade secret information, such as new product information or manufacturing process details, could be extremely valuable to any chemical firm. The proprietary TSCA CDS, with its large amount of trade secret data, therefore should be considered a prime target for wiretapping threats.

4.3 Computer Technology Security Mechanisms and Techniques

4.3.1 Introduction

Four basic methods of providing security with computer technology exist:

- Access Control
- Surveillance
- Integrity Monitoring
- Encryption

These methods are utilized to protect the computer systems and its contents against the security threats discussed in Section 4.2. They are implemented in a variety of software and hardware combinations.

The following discussion examines the security methods, their various application possibilities, and the threats they are designed to counter. Specific applications appropriate for the TSCA CDS will be identified, and software and hardware capabilities required for implementation will be discussed in the final report.

4.3.2 Access Control

Access control is the primary method for providing computer system security in a data sharing environment. Its basis is the principle of least privilege[3]; i.e., users are authorized to access only the data and system resources which they have a legitimate need to use. This principle means that users are limited in number and in the specific resources they can use. Users are allowed access only after they are granted explicit permission, which is based on a user's "need-to-know". Any unauthorized access attempt should result in denial of user access.

Limiting the number of users and the extent of their privileges provides security control because it reduces opportunities for deliberate or accidental misuse of the system. It restricts the ability of outsiders to gain physical access to authorized terminals, thus reducing their capabilities to penetrate system safeguards. Restrictions should result in the fewest number of people knowing or having access to the least amount of sensitive or confidential information possible. The amount of confidential information needed, of course, will depend on an individual's job responsibilities. Restrictions also enhance accountability for security violations by limiting the number of possible suspects. Individual accountability is a strong deterrent to deliberate system misuse by authorized users.

---

[3] Jerome H. Saltzer and Michael D. Schroeder, "The Protection of Information in Computer Systems", in Proceedings of the IEEE, Volume 63, Number 9 (September 1975), p. 1282.

Access control methods mechanize and enforce administrative security policies through authorization definition and checking. Authorization definition supplies the following information to the system:

- A set of valid system users
- Information necessary to verify user identity uniquely
- A set of system resources
- Data needed to verify identity of the resources
- Valid authorization relationships between users and resources[4]

The first four items concern identification/authentication information, which is used to define the valid authorization relationships for system use. Authorization checking in turn uses these relationships to ensure that users have the right to access the system resources they have requested.

Positive identification of system users and resources is a key requirement for computer system security because it provides the reference points by which authorization checking determines who has access to what resources. A detailed positive identification involves determining each user, computer program, group of data, and peripheral worthy of protection. Each one is assigned a unique identifier and is registered with the system.

---

[4] I. S. Broadman, "Protection Techniques in Data Processing Systems to Meet User Data Security Needs", in Computer Communication Today and up to 1985. Second International Conference on Computer Communication, Gaithersburg, Maryland, 1974, pp. 486-487.

A detailed identification scheme permits greater accountability of system activities and more stringent access control specifications. For example, each individual user should be identified in order to promote individual accountability for security threats. By the principle of least privilege, only the data elements needed for a specific operation should be accessible; thus each data element should be identified separately in order to expose the data to the least possible risk. Programs and peripherals should be identified in order to maintain precise control of system resources.

Disadvantages of a detailed identification scheme are the large amount of overhead required to process the detailed information and to perform extensive authorization checking.

The unique identifiers enable each user and resource to be identified to the system during operations. Resources under the physical control of the computer, such as disk packs or input-output channels, can be identified positively by the computer. When a user submits an identifier, positive identification cannot be assured. Thus, authentication must occur to complete the system's identification process.

Authentication mechanisms verify a user's claimed identity. Verification of true identity depends on the user submitting a secret password or an unforgeable object. In a typical access attempt, the user enters a name into the system and supplies a password in response to a system request. The submitted password is compared with the password the system expects from that user. If the passwords match, the user's identity is verified and log on procedures can be completed. Verification by password is the most common method of authentication but it has disadvantages because:

- user-supplied passwords usually are short and are remarkably easy to guess
- hard-to-remember passwords invite written lists
- group passwords encourage exchanging passwords among users
- passwords must be exposed to perform their functions
- password lists are vulnerable to exposure

To counter these difficulties, passwords should be of sufficient length to thwart systematic, exhaustive guessing and to reduce the probability of success by random guessing. They should be system-generated sequences of random characters. To enhance user memorization and discourage written records, passwords should be made pronounceable by the use of digraph statistics or similar techniques. Borrowing or exchanging passwords can be discouraged by assigning a unique password to each user. Exposure of passwords during use can be prevented through the use of over printing or print suppression features on terminals. Encryption can protect passwords transmitted over communications lines. The password list itself is of major security importance and should be protected at the highest possible level. One such protection method uses a hard-to-invert transform to encrypt the password list. Possession of the encrypted passwords reveals nothing about the passwords input by users.

Passwords should be changed immediately any time exposure is known or suspected. Also, passwords should be changed with varying frequency as a preventive measure.

Identity authentication can be accomplished by requiring a user to submit a unique and unforgeable item to the system. Such verification by user fingerprints, voiceprints, or keystroke style is in the research stages currently and is not yet practical for operational systems.

Magnetized plastic cards are a practical and effective alternative. A user inserts his/her card into a terminal which reads the magnetic code and transmits the information to the computer. The authentication mechanism compares the received code with its expectation from that user and verifies the user's identity. The weaknesses of using magnetized cards are that they may be lost, stolen, or counterfeited and used by intruders. They are subject to damage or erasure of the code from proximity to magnets or other magnetized cards. Good physical and personnel security methods should minimize these risks.

Authentication techniques based on passwords and/or magnetized cards are subject to interception by wiretapping. Encryption of the data streams before they are transmitted from remote terminals alleivate this risk. One method combines the authentication and encryption processes. A user enters a name, then loads a designated transformation key. When the computer receives the user's name, it loads its corresponding encryption key. Properly matched encryption keys allow successful communication between user and system, thus signalling that authentication has occurred.

In authorization definition, users and resources registered during the identification process are distinguished as subjects and objects.[2] Subjects are active system agents or processes that make requests for data or resources. Examples of subjects are:

- Users
- Terminals
- Programs

Objects are passive system agents or processes that are requested by subjects. Examples include:

- Programs
- Data
- Peripherals

Some items can be both subjects and objects, depending on their current operational functions. For example, programs make requests for data, and also are requested themselves.

Based on administrative access authorization decisions, each subject is associated with the objects it is authorized to access. Association of subjects and objects in valid relationships is accomplished in two ways:

- by listing each subject with all the objects it is authorized to access; or,
- by listing each object with all the subjects authorized to access it.

Applying additional restrictions to the valid relationships achieves further access control. Execution privileges can be controlled so that subjects have read, write, or execute privileges, or any combination of them. There is much less chance of data manipulation or damage if a user is allowed only to read a file rather than writing into it. Objects also can have execution restrictions associated with them. For example, users should not be allowed to read, write, or execute the software that provides protection.

---

[5] Steven B. Lipner, "Secure Computer Systems for Network Applications", in Network Structures in an Evolving Operational Environment, IEEE Fourth Data Communications Symposium, October 7-9, 1975, p. 8-9.

Authorization checking is the process that determines whether an attempted access can be achieved. Access is permitted only if the subject-object relationships of the attempt match the previously defined valid subject-object relationship. The process basically mediates between subjects and objects. Before a subject can complete an access, the authorization checking process determines if it has the right to make that access.

In conjunction with a detailed identification scheme, mediation can provide stringent security. In a totally secure system, "the primary underpinning of the protection system" is the principle that "every access to every object must be checked for authority"[6]. This means that all objects and subjects are under control of the system. Such detailed mediation is infrequent in practice because of the huge processing overhead involved. Most systems provide selective mediation for security-related access attempts.

Authorization checking methods are implemented in either list-oriented applications, ticket-oriented applications, or a combination of both. List-oriented mechanisms are organized around lists of subjects with the objects they are authorized to access, or lists of objects with the subjects that are authorized to access them. When a user requests access to certain data, the authorization checking mechanism searches that user's list to determine if he/she has been granted access privileges to that data. Alternatively, the checking mechanism is designed to search the list associated with the protected data to determine if the user's identifier is included. List-oriented applications provide clear accountability for system activities because all access attempts are based on identification of subject and object. They also allow easy identification of system privileges granted to each subject. The main disadvantage of list-oriented mechanisms is that they require associative searching and this incur added processing overhead. If a detailed idenitification scheme is used in conjunction with access control lists, the processing inefficiency may be intolerable.

Ticket-oriented applications assign each object a unique key, then give appropriate tickets to subjects that have been granted access authorization. When a user makes an access request, he/she presents a ticket to the mediation mechanism. That ticket is compared with the object's assigned key and if they match access is granted.

---

[6] Saltzer and Schroeder, "The Protection of Information in Computer Systems", p. 1282.

Because the system makes only a comparison, ticket-oriented mechanisms are much faster than list-oriented ones. The former do not allow accountability for system activities since any user with a proper ticket can access an object. Tickets are subject to the same weaknesses as passwords. Ticket-oriented mechanisms also place the burden on the user to determine what ticket to present for any access.

To overcome the difficulties inherent in each type of mechanism, some systems use both. List-oriented activities are used at the user interface and for security-related authorization checking. Ticket-oriented implementations are used to provide detailed access mediation.

4.3.3 Surveillance

Computer surveillance is defined as the process of monitoring and recording the movement of programs and data within the system. Commonly, there are two types of security surveillance: auditing and threat monitoring. Auditing, which includes logging and journaling, involves continual recording of a set of facts about user requests and their system responses. Audit trails can be generated for any desired level of detail. For security purposes, a minimum audit trail should consist of a small set of facts concerning an access request and corresponding system decisions associated with the access request. The audit should supply sufficient data to allow systems security personnel or external auditors to analyze the information. Analysis can uncover possible system misuse and identify the responsible party, and can indicate vulnerabilities which may require increased protection.

One approach to identifying misuse is to log the events associated with threat monitoring activities. These events include:

- Jobs on and off, to include user/terminal/program identification, date and time

- Data requested

- Files accessed, to include type, number of accesses and access keys

- Disposition of data (number of records output or displayed, program status information, etc.)

- Unauthorized log-ons
- Response to random queries
- Any special use of the system
- All descriptor changes
- Configuration changes
- Changes to security tables and the supervisor
- Restarts and machine faults
- Attempted violations of memory
- Aborts and parity errors [7]

Another solution is to record all data movement or to record all facts necessary to reconstruct an affected data base. Although misuse of the system can be treated similarly to a system failure, misuse is difficult to locate because of the large volume of data collected.

A potential solution is to establish two logs: one for failure recovery, and one for security. The security audit would record rejections of attempted access, thus clearly identifying any attempted misuse of the system or data. The recovery audit then would be examined in further detail to trace all the data movements.

Threat monitoring is a surveillance technique that involves collecting information on the operation of the security system in a real time mode in order to detect unauthorized, invalid, or unusual activities. The results can be handled by trap or interrupt procedures that take appropriate action, depending on the perceived (or anticipated) severity of the threat. The user may be allowed a minimal number of invalid access attempts before he/she is disconnected, and/or the user may be required to begin at a previous valid procedure, or the user session may be terminated immediately. In almost all cases, as soon as any threatening activities are detected, the threat monitoring system switches control of the system from the user state to the privileged or supervisor state which then can require additional user authorization before allowing completion of the attempted action.

---

[7] Peter S. Browne, "Taxonomy of Security and Integrity", in Security and Privacy in Computer Systems, ed. Lance J. Hoffman (Los Angeles, Calif., Melville Publishing Co., 1973), p. 378.

In addition to detecting and responding to unauthorized, invalid, or unusual activities, threat monitoring includes recording the activities and system responses. The threat monitoring log should include the information to allow the threat monitoring procedure or the operating system to determine the kind and extent of the threatening activity and to take the appropriate action.

It is important to note that both the auditing and threat monitoring logs and the threat monitoring procedure are themselves confidential information and should be protected. Exposure of any of them, particularly the threat monitoring procedure, could lead to subversion of system security.

4.3.4 Integrity Monitoring

Integrity monitoring insures that security methods enforce security policies. Integrity monitoring is similar to surveillance, particularly threat monitoring, although it usually is concerned primarily with making appropriate real time responses to threats. Integrity monitoring also includes preventive security measures as well as corrective ones.

Integrity monitoring considerations apply to two areas, data integrity and system integrity. Data integrity monitoring provides protection against:

- Damage
- Update Conflict
- Processing of Incorrect Data

Examples of integrity monitoring techniques include:

- Maintenance of system, data, and program backups for system recovery

- Validity checking of data transfers

- Validity checking of data by reasonableness checks, consistency checks, range checks, and sequence of events checks

- Audit trail of program modules

- Lockout for protection of recently modified data

- Validation required for on-line updates[8]

---

[8] Browne, "Taxonomy of Security and Integrity", p. 376.

Integrity monitoring includes both hardware and software components. In most cases, hardware is monitored mainly to safeguard the system against damage from hardware failures. The integrity monitoring procedure takes appropriate action to protect the system's resources against damage and exposure. However, most operating systems are vulnerable to hardware failure, particularly failures resulting in undetected disability of protection mechanisms. Some systems may require implementation of hardware threat monitoring mechanisms. System crashes caused by software failures are handled the same way as hardware failures.

Integrity monitoring frequently must identify and respond to problems caused by unreliable system and user software. Powerful debugging and testing tools are used to improve software reliability, especially for system software. Stringent security requires all system software to be audited and verified, particularly the software for sensitive applications. Security-related code should be audited line-by-line in any security environment.

Integrity monitoring protects the computer system from threats other than failures. Integrity monitoring procedures determine the nature and extent of any threats, and then take appropriate action in real time to protect the system and data. For example, monitors may limit the number of repeated invalid attempts to access, and may cause the system to shut down immediately as the result of some types of threats. Integrity monitoring can also take preventive actions such as continuous or sporadic authentication. Another preventive action is that user-requested address spaces can be checked for validity before access is allowed. In these ways, integrity monitoring enforces access control mechanisms and policies.

4.3.5 Encryption

Encryption is the process of transforming data representation so that its information content is masked from anyone who does not have the inverse transform. It is used primarily for transmission of confidential data over insecure communications lines to thwart wire tapping threats; but it can also be utilized effectively for protecting stored data from exposure to unauthorized users as well.

Most encryption schemes consist of two parts, a transformation algorithm and a secret key. The algorithm indicates the procedures and computations that are performed in order to scramble the data. The secret key supplies a value so that data stream values are changed in a specific way. Data then can be decoded only by possession and application of the inverse transform loaded with the same secret key.

In order to effectively protect data, an encryption scheme must be difficult to cryptanalyze. The potential value of trade-scret data in the proprietary TSCA CDS will require an encryption scheme strongly resistant to crytanalysis for any data transfer over unsecured communcations lines. Resistance to cryptanalysis is achieved by maintaining the secrecy of the transformation process. In many cases, encryption algorithms are publicly known, so secrecy depends on a combination of the algorithm's sophistication and protection of the key. Since the key provides the crucial value determining the encryption, its exposure could compromise system security by allowing wire tappers to break the code. Therefore, the key must be kept strictly secret from all except authorized users, and it must be changed whenever any potential exposure occurs.

The key also should be difficult to guess or infer from analysis of encrypted data. The algorithm should allow for a large number of valid keys so that any specific key cannot be guessed at random or deduced by elimination. Furthermore, outsiders should not be able to deduce the key by inference from the representation of the data. Thus encryption involves obscuring the data patterns as well as changing the data values. Transformations on a character-by-character and word-by-word level are especially vulnerable to detection because they provide information content about the data. Two alternative encryption approaches obscure the natural language structure; one encrypts according to a standard, preset number of characters, while the other encrypts in a continuous stream, with no spaces appearing between transformed characters.

In addition to security over communications lines, encryption can protect data, such as confidential information, passwords, and access lists, that are stored in the system. Although most such schemes are similar to those described above, one password application uses an algorithm that is extremely difficult to invert, thereby thwarting attempts to deduce general password structure.

Encryption of stored data is extremely effective as a protection mechanism, although limitations exist[9]. Due to the nature of the transformation process, it is necessary to process information in clear text, thus exposing it to threats. It also does not protect against data destruction, either in storage or transmission. In an interactive system, the cipher keys must be stored on-line. To avoid compromise, encryption information should be protected at the highest security level.

Encryption is insufficient alone as a security method because of a system's processing requirements, physical data protection requirements, and cipher key protection requirements. It can enhance security and is essential for transmitting confidential data over unsecured communications lines.

4.3.6 Software

Software associated with computer system security include programs written for access control, surveillance, integrity monitoring, and encryption, as well as application programs used to process the data base.

In order for software to provide the maximum amount of security, it should be developed in-house. There are, however, many commercially available packages that perform the security functions described in the previous four sections. If commercial packages are used, they should be audited to determine any potential security risks. This could be done by developing comprehensive test programs and procedures to verify the protection features of all operations performed by the package. The disadvantages of packages are that most systems have unique security needs resulting in possible modifications to the package, and procedures are generalized resulting in increased computer processing time.

Software developed in-house should be modularly designed. Programming responsibilities for various modules should be designated so that no single programmer will know of the security-oriented features of the system. Higher-level programming languages (i.e., COBOL, PL/1) should be utilized as much as possible. Assembly language programming is more difficult to audit and provides an easier mechanism to switch from a user state to a supervisor state. All software should be audited on a line-by-line basis to verify that it performs the required function and does not contain extraneous coding that might enable security to be bypassed. Programs should be protected by a system of passwords for execution and modification privileges. Program modifications should be controlled automatically by means of an audit trail that logs and verifies all changes.

---

[9] Gerald J. Popek, "On the Technical State of the Art in Computer Security", in Eleventh IEEE Computer Society International Conference (Compcon), 1975, p. 41.

A subverter program should be developed to test the security of the system. This program would generate false identifications, unauthorized attempts to access application programs, data, or the Supervisory Program, simulated system failures, and other probes to test essential features of the security system.

4.3.7 Hardware

The hardware and its associated Supervisory Program (operating system) are principle sources of the systems security features. Since the Supervisory Program controls user access to itself as well as user information and computer programs, it provides a means of protection from intrusion. However, a lack of adequate security features in most Supervisory Programs has made them prime targets for unauthorized system access or penetration. System intruders employ various means to bypass, penetrate, or subvert the Supervisory Program's protection mechanism. Once in a privileged state, intruders can access or change any system resource unit or area of core storage.

Security features that pertain to an on-line, multi-user, computer system such as TSCA CDS are discussed for the following hardware-related categories:

- Central Processor
- Main Storage
- I/O Channels
- Front End Processors
- Terminals
- Secondary Storage Devices

An important feature of the central processor is the ability to prevent intruders from interferring with the processor's control as well as the control of other system components. This is accomplished by preventing programs from executing certain processor instructions through incorporation of at least two execution states within the processor; i.e., a privileged (supervisor) and a user state. The privileged state consists of instructions for executing system resource control functions not available in the user state. Switching from the user state to the privileged state is accomplished by executing special instructions that require authorization checks every time they are executed. Checks can also be made to insure that certain instructions are executed in a prescribed sequence.

Another security feature of the central processor is the ability to detect and respond to various types of system malfunctions and improper operations through the use of interrupt processors. These include attempts to:

- Execute illegal, invalid, or privileged instructions
- Circumvent core storage protection features
- Execute programs which cause the processor to loop or time-out

The interrupt processor is built into the central processor and receives control whenever a malfunction is detected. It can then log the security violation, request additional authorization information or terminate the computer program.

The central processor should also provide hardware-implemented read, write, and execute privileges. Every time a user attempts to exercise one of these privileges, it is automatically checked by the hardware. Unauthorized attempts produce an interrupt that returns control to the supervisor state which will take the appropriate action.

Main storage security features include protection mechanisms for preventing users from accessing unauthorized information in the address space of other users or the Supervisory Program. The goal is to isolate users from each other. In a non-virtual storage system this is accomplished by core storage address boundaries or protection keys. In a virtual storage system a descriptor-based protection mechanism can be employed. In this type of system storage, areas are divided into segments which are uniquely identified and addressed. Accessing capabilities are assigned to each segment. All memory accesses must go through both an addressing check and a descriptor check before execution takes place.

I/O channels and controllers provide the interface between main storage and peripheral devices. I/O storage control mechanisms assure that a data transfer is made to the storage blocks specified by the I/O command and that the storage block is within the address space allocated to the user program.

Front-end processors, which are usually minicomputers that function as communications controllers, require mechanisms to detect errors. When there is "noise" in the line and it is disconnected, another user may get the line before the host computer is notified of the disconnection resulting in information given to the wrong person. This situation can be averted by employing adequate interrupt processors.

Security for interactive terminals is accomplished by "hard-wiring" the identification of the terminal in the Supervisory Program. This will allow only a particular computer program and a particular terminal to transfer information to each other. The user could access a computer program that retrieved confidential information only by using a particular terminal that would be in a secure area. Terminals can also be wired for update only or retrieval only.

Secondary storage consists of disks and magnetic tapes. Security mechanisms include read-only locks that preclude that device from being written into by any program, including the Supervisory Program. Hardware can also provide erase mechanisms to purge sensitive data from the device.

LIST OF WORKS CONSULTED

Abbott, R. P., J. S. Chin, J. E. Donnelley, et al. *Security Analysis and Enhancements of Computer Operating Systems.* Springfield, Va.: National Technical Information Service, 1976. (PB 257 087)

Attanasio, C. R., P. W. Markstein, and R. J. Phillips. "Penetrating an Operating System: A Study of VM/370 Integrity." *IBM Systems Journal*, Vol. 15, No. 1 (1976). pp. 102-116

Bjork, Lawrence A., Jr. "Generalized Audit Trail Requirements and Concepts for Data Base Applications." *IBM Systems Journal*, Vol. 14, No. 3 (1975). pp. 229-245

Broadman, I. S. "Protection Techniques in Data Processing Systems to Meet User Data Security Needs." *Computer Communication Today and up to 1985.* Second International Conference on Computer Communication, Gaithersburg, Md.: 1974. pp. 485-489

Browne, Peter S. "Taxonomy of Security and Integrity." *Security and Privacy in Computer Systems*, ed. Lance J. Hoffman. Los Angeles, Calif.: Melville Publishing Company, 1973. pp. 369-378

Burke, E. L., 1Lt. W. Price, and J. C. C. White. "Secure Minicomputer Systems Architecture." *Micros and Minis: Application and Design*, Ninth Annual IEEE Computer Society International Conference (Compcon), Washington, D. C.: 1974. pp. 95-97

Bushkin, Arthur A. *A Framework for Computer Security.* Revised edition. Springfield, Va.: National Technical Information Service, 1975. (AD-A025 356)

Controlled Accessibility Workshop Report. National Bureau of Standards Technical Note 827. Washington, D. C.: U.S. Department of Commerce, National Bureau of Standards, 1974.

Conway, R., W. Maxwell, and H. Morgan. "On the Implementation of Security Measures in Information Systems." *Communications of the ACM,* Vol. 15, No. 4 (April 1972). pp. 211-220

Donovan, J. J., and S. E. Madnick. "Hierarchical Approach to Computer System Integrity." *IBM Systems Journal*, Vol. 14, No. 2 (1975). pp. 188-202

Hinke, Thomas H., and Marvin Schaefer. *Secure Data Management System.* Springfield, Va.: National Technical Information Service, 1975. (AD-A019 201)

Hoffman, Lance J., ed. *Security and Privacy in Computer Systems.* Los Angeles, Calif.: Melville Publishing Company, 1973.

Linden, Theodore A. *Operating System Structures to Support Security and Reliable Software.* National Bureau of Standards Technical Note 919. Washington, D. C.: U.S. Department of Commerce, National Bureau of Standards, 1976. (Also available from National Technical Information Service, PB 257 421.)

Lipner, Steven B. "Secure Computer Systems for Network Application." *Network Structures in an Evolving Operational Environment*, IEEE Fourth Data Communications Symposium, 7-9 October 1975. pp. 8-8-8-12.

Lowe, Thomas C. *Computer Security Guidelines for Implementing the Privacy Act of 1974.* Federation of Information Processing Societies Publication 41. Washington, D. C.: U.S. Government Printing Office, 1975.

Popek, Gerald J. "On the Technical State of the Art in Computer Security." *Eleventh IEEE Computer Society International Conference* (Compcon), 1975. pp. 40-41

Popek, Gerald J., and Charles S. Kline. "A Verificable Protection System." *Proceedings of the 1975 International Conference on Reliable Software*, 21-23 April 1975, Los Angeles, Calif. pp. 294-304.

Renninger, Clark R., ed. *Approaches to Privacy and Security in Computer Systems* National Bureau of Standards Special Publication 404. Washington, D. C.: U.S. Department of Commerce, National Bureau of Standards, 1974.

Saltzer, Jerome H. "Protection and the Control of Information Sharing in Multics." *Communications of the ACM,* Vol. 17, No. 7 (July 1974). pp. 388-402.

Saltzer, Jerome H., and Michael D. Schroeder. "The Protection of Information in Computer Systems." *Proceedings of the IEEE*, Vol. 63, No. 9 (Sept. 1975). pp. 1278-1308.

Schroeder, Michael D., and Jerome H. Saltzer. "A Hardware Architecture for Implementing Protection Rings." *Communications of the ACM*, Vol. 15, No. 3 (March 1972). pp. 157-170.

Smith, L. *Architectures for Secure Computing Systems.* Springfield, Va.: National Technical Information Service, 1975. (AD-A009 221)

Turn, Rein. *Cost Implications of Privacy Protection in Databank Systems.* Springfield, Va.: National Technical Information Service, 1975. (AD-A022 186)

Turn, Rein, and Willis H. Ware. *Privacy and Security in Computer Systems.* Springfield, Va.: National Technical Information Service, 1975. (AD-A016 493)

White, Richard. *Computer Security Technology Reference Manual.* Springfield, Va.: National Technical Information Service, 1975. (AD-A019 439)

5.0 EPA/OTS REQUIREMENTS RELATED TO THE CHEMICAL SUBSTANCES INFORMATION NETWORK

5.1 Introduction

The primary objective of the proposed Chemical Substances Information Network (CSIN) is to provide information to those federal agencies, commercial enterprises, and the general public involved in the study and regulation of chemical substances. The CSIN would consist of data bases and systems distributed throughout a network. The network would be developed in an evolutionary manner depending on the needs and the availability of funds of particular users. The CSIN would be developed by building on existing systems, implementing new systems, and combining several data bases into a single data base system.

The Toxic Substances Control Act will require EPA, through the Office of Toxic Substances, to collect and have available for its use a vast amount of information pertaining to chemical substances. Under TSCA, OTS will be required to conduct risk assessments, predict toxicity, select chemicals for testing, collect information on chemical substances, approve chemicals for manufacturing, create an inventory of chemical substances, and regulate new and existing chemical substances. Given these requirements, EPA/OTS will be a major user and developer of data bases within the Chemical Substances Information System and will play a prime role in its design and implementation.

5.1.1 Background

The Mitre report identified the data bases and systems that would be included in the Chemical Substances Information Network. The data bases were divided into two groups: core components and non-core components. The core components consist of data bases that are of prime importance to OTS in carrying out its requirements under TSCA. The Mitre report recommends that the core components be developed using a common standardized nomenclature, be maintained under a single data base management system to facilitate cross exchanges of information, and be directly linked wherever possible to provide rapid access to needed information. The non-core components provide auxilliary information that would also be important to OTS. Linkage to these data bases would be by referral that would be provided through the Chemical Data Bases Directory.

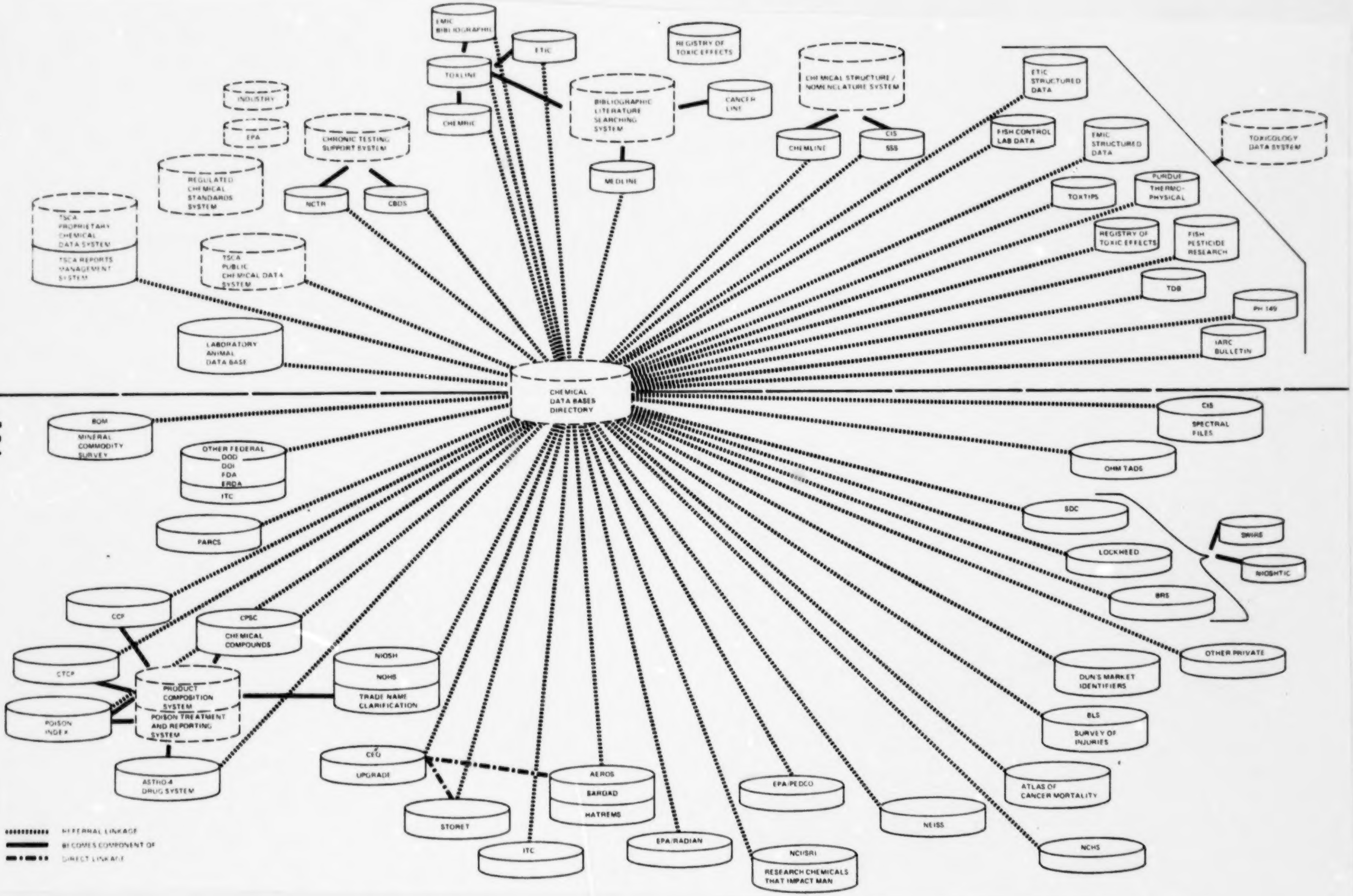
Figure 1-1 in the introduction of this report illustrates the long-term implementation objectives of the Chemical Substances Information Network. It indicates data bases that do not currently exist and must be developed, and data bases that were developed by integrating several data bases that currently exist. Figure 5-1, also taken from the Mitre report, illustrates CSIN components as they currently exist. It indicates the data bases that must be developed, and indicates individual data bases that would be integrated into various long-term data base systems illustrated in Figure 1-1.

5.1.2 Approach

The Mitre report developed a time phased implementation plan for each of the ten core components of the CSIN. The plan showed how the existing data bases illustrated in Figure 5-1 would evolve into the ten core components illustrated in Figure 1-1 over a ten year period. Two of these data base systems, the Chemical Data Bases Directory and the Chemical Structure/Nomenclature System, have been discussed in detail in Sections 2.0 and 3.0. Five other data base systems, which included the Toxicology Data System, the Bibliographic Literature Search System, the Chronic Testing Support System, the Regulated Chemicals Standards System, and the Laboratory Animal Data Base, require additional detailed feasibility studies over the next few years before effective implementation recommendations could be made. Sufficient user requirements and needs for these five data base systems are currently undefined, and therefore were not considered within the scope of this report. The three other data base systems, the proprietary TSCA CDS, the public TSCA CDS, and the TSCA Reports Management System, are directly related to the reporting requirements under TSCA, and therefore are of prime importance to the operations of the Office of Toxic Substances. Implementation recommendations for these three systems are combined and discussed under the TSCA Chemical Data System. Information for these recommendations was obtained from EPA strategy documents concerning TSCA, TSCA requirements published in the Federal Register, and interviews with appropriate OTS personnel. In order to discuss various OTS interactions with the Chemical Substances Informations Network, Sigma Data identified several functional groups within OTS. The activities of each functional group are discussed as they relate to other functional groups and CSIN component data base systems. Finally, short-term and long-term network configurations are presented showing how OTS would be linked to the core components of the Chemical Substances Information Network.

FIGURE 5-1

EXISTING CHEMICAL SUBSTANCES INFORMATION NETWORK


CHEMICAL DATA BASES DIRECTORY
EMIC BIBLIOGRAPHIC
ETIC
TOXLINE
CHEMRIC
REGISTRY OF TOXIC EFFECTS
BIBLIOGRAPHIC LITERATURE SEARCHING SYSTEM
CANCER LINE
MEDLINE
CHEMICAL STRUCTURE / NOMENCLATURE SYSTEM
CHEMLINE
CIS SSS
INDUSTRY
EPA
CHRONIC TESTING SUPPORT SYSTEM
NCTR
CBDS
REGULATED CHEMICAL STANDARDS SYSTEM
TSCA PROPRIETARY CHEMICAL DATA SYSTEM
TSCA REPORTS MANAGEMENT SYSTEM
TSCA PUBLIC CHEMICAL DATA SYSTEM
LABORATORY ANIMAL DATA BASE
ETIC STRUCTURED DATA
FISH CONTROL LAB DATA
EMIC STRUCTURED DATA
TOXICOLOGY DATA SYSTEM
TOXTIPS
PURDUE THERMO-PHYSICAL
REGISTRY OF TOXIC EFFECTS
FISH PESTICIDE RESEARCH
TDB
PH 149
IARC BULLETIN
BOM MINERAL COMMODITY SURVEY
OTHER FEDERAL
DOD
DOI
FDA
ERDA
ITC
PARCS
CCP
CPSC CHEMICAL COMPOUNDS
CTCP
PRODUCT COMPOSITION SYSTEM
POISON TREATMENT AND REPORTING SYSTEM
POISON INDEX
ASTHO-4 DRUG SYSTEM
NIOSH
NOHS
TRADE NAME CLARIFICATION
CEQ
UPGRADE
STORET
ITC
AEROS
SAROAD
HATREMS
EPA/RADIAN
NCI/SRI RESEARCH CHEMICALS THAT IMPACT MAN
EPA/PEDCO
NEISS
NCHS
ATLAS OF CANCER MORTALITY
BLS SURVEY OF INJURIES
DUN'S MARKET IDENTIFIERS
OTHER PRIVATE
CIS SPECTRAL FILES
OHM TADS
SDC
LOCKHEED
BRS
SWIRS
NIOSHTIC
REFERRAL LINKAGE
BECOMES COMPONENT OF
DIRECT LINKAGE

5.2 TSCA Chemical Data System

The TSCA Chemical Data System would be the major data base system within OTS to hold information submitted as a result of TSCA. The data base would contain extensive information on chemicals in commerce and commercial sites that manufacture or process these chemicals. The TSCA CDS would consist of two data bases: a proprietary TSCA CDS and a public TSCA CDS. The proprietary data base would contain both confidential and non-confidential information and would only be accessed by OTS personnel. Stringent security requirements would be imposed on this data base. The public data base would be a subset of the proprietary data base and only contain non-confidential information. This data base would be available for general use and have no security requirements. Sigma Data recommends combining the Reports Management System with the proprietary TSCA CDS under a single data base management system. TSCA CDS requirements are described in more detail through discussions concerning information flow, data base organization and structure, storage estimates and search and retrieval capabilities. Based on the requirements two alternative implementation plans are presented.

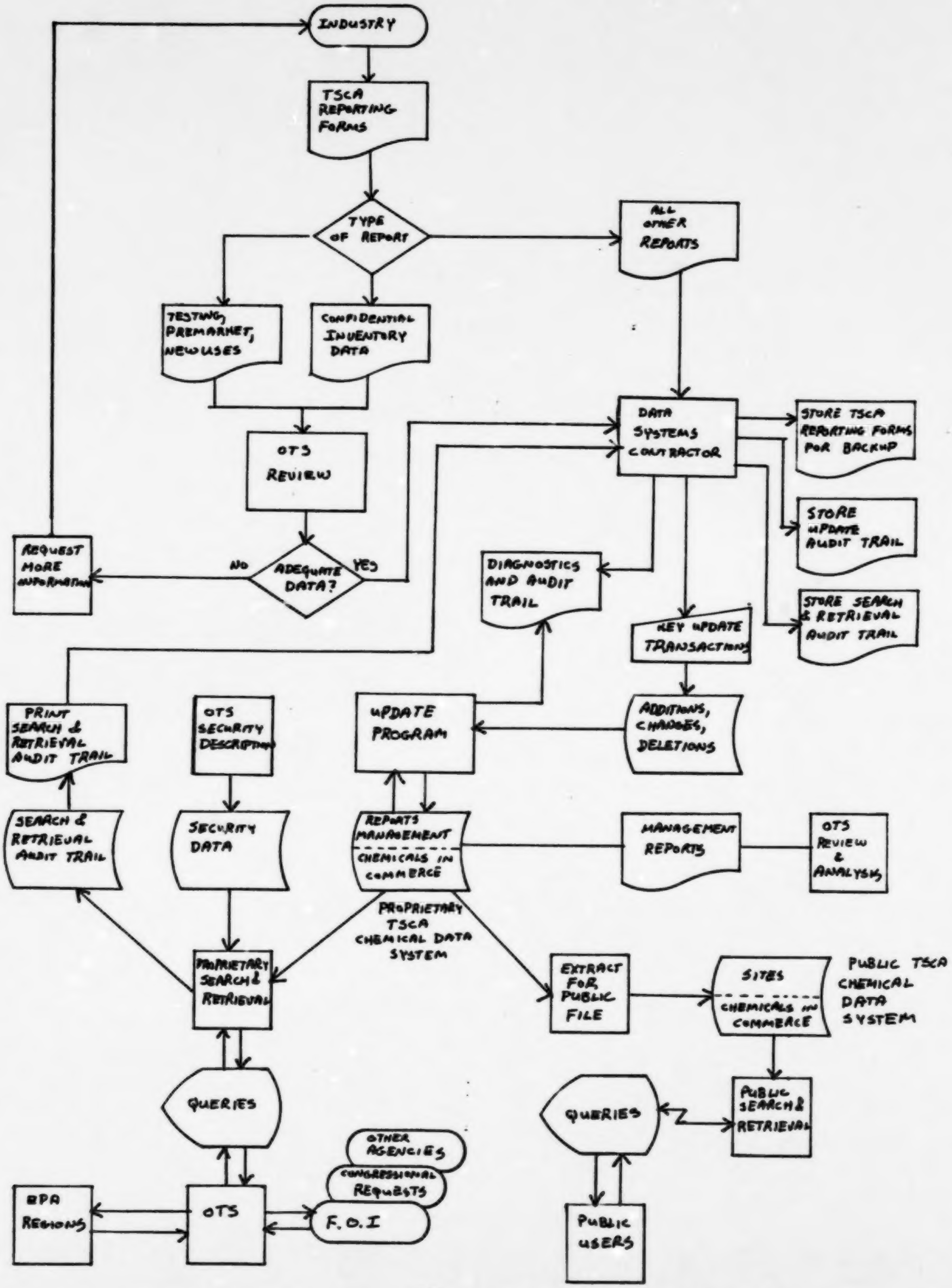
5.2.1 General System Description

This section describes, in general terms, the functional characteristics of a chemical data system that could be used by OTS to collect, store, and disseminate information submitted by industry in response to TSCA reporting requirements. The proposed system would be a combination of three core components described in the Mitre report; the proprietary TSCA Chemical Data System, the Public TSCA Chemical Data System, and the Reports Management System. Figure 5-2 illustrates the flow of information through the proposed TSCA Chemical Data System (CDS). The diagram would be applicable to each of the various stages of TSCA reporting that OTS might implement.

The information submitted to EPA by industry would be processed and stored in a data base known as the Proprietary TSCA Chemical Data System. The data base is divided into two files, Reports Management and Chemicals in Commerce, each of which has a hierarchical data structure. The first file provides a tracking capability for TSCA respondents. It is organized according to sites where chemicals are manufactured, processed or imported, and contains site identification information. Associated with each site is information pertaining to the identification of reports submitted by a particular respondent. The reports would include site specific production, premarket notification, uses, exposure, byproducts, impurities, adverse reactions, health data, test data, new uses etc.

FIGURE 5-2

TSCA CHEMICAL DATA SYSTEM INFORMATION FLOW DIAGRAM


INDUSTRY
TSCA REPORTING FORMS
TYPE OF REPORT
ALL OTHER REPORTS
TESTING, PREMARKET, NEW USES
CONFIDENTIAL INVENTORY DATA
OTS REVIEW
DATA SYSTEMS CONTRACTOR
STORE TSCA REPORTING FORMS FOR BACKUP
STORE UPDATE AUDIT TRAIL
STORE SEARCH & RETRIEVAL AUDIT TRAIL
REQUEST MORE INFORMATION
NO
ADEQUATE DATA?
YES
DIAGNOSTICS AND AUDIT TRAIL
KEY UPDATE TRANSACTIONS
PRINT SEARCH & RETRIEVAL AUDIT TRAIL
OTS SECURITY DESCRIPTION
UPDATE PROGRAM
ADDITIONS, CHANGES, DELETIONS
SEARCH & RETRIEVAL AUDIT TRAIL
SECURITY DATA
REPORTS MANAGEMENT
CHEMICALS IN COMMERCE
MANAGEMENT REPORTS
OTS REVIEW & ANALYSIS
PROPRIETARY TSCA CHEMICAL DATA SYSTEM
PROPRIETARY SEARCH & RETRIEVAL
EXTRACT FOR PUBLIC FILE
SITES
CHEMICALS IN COMMERCE
PUBLIC TSCA CHEMICAL DATA SYSTEM
QUERIES
PUBLIC SEARCH & RETRIEVAL
QUERIES
OTHER AGENCIES
CONGRESSIONAL REQUESTS
EPA REGIONS
OTS
F.O.I
PUBLIC USERS

The second file contains the data submitted on various reports and is organized according to chemical compound identification characteristics. Associated with each chemical is information pertaining to production, industrial use, occupational exposure, adverse reactions, health data, byproducts and impurities. The actual information at any one time will depend on the TSCA reporting requirements implemented by OTS. The two segments are processed under a single data base management system to provide a direct link between a chemical compound and a particular site where it is manufactured, processed, or imported. Section 5.2.5 discusses the data base organization and structure in more detail.

All reports submitted by industry as a result of TSCA reporting requirements would be sent to the Office of Toxic Substances at EPA. Reports that contain confidential information would be carefully reviewed by OTS for completeness and correctness. Reports containing missing or invalid data would be returned to the submitter for correction. Confidential reports that pass the OTS review and non-confidential reports are forwarded to the data systems contractor for entry into the TSCA CDS.

The data systems contractor would maintain a log of all reports submitted and store all the reports as a backup procedure. The contractor would convert the information on the reporting forms to a set of transactions to be processed by an update program. The update program would apply the transactions to the TSCA CDS and produce a diagnostic listing that would print processing statistics and indicate any reporting errors that were detected during the update process. The contractor would analyze the diagnostic listing and return those reports containing errors to the submitter for correction. The diagnostic listing would also be stored by the data systems contractor to provide an audit trail of information entered into the TSCA CDS.

The proprietary TSCA CDS would contain both confidential and non-confidential information and would be accessed on-line exclusively by the Office of Toxic Substances. However, EPA also has an obligation to make as much of this information as possible available to the public user while still protecting the confidentiality claims submitted by industry. In order to accomplish this, a separate public TSCA CDS would be constructed from the information found in the proprietary TSCA CDS, as suggested in the Mitre report. Each data element in the proprietary TSCA CDS that can be declared as confidential will have an associated data element indicating the type of confidentiality that is actually claimed. An extract program would process the entire proprietary TSCA CDS to extract non-confidential information. By analyzing the various confidentiality codes in an hierarchial manner, the program would produce a public TSCA CDS with a similar data base organization and structure, but delete or disguise any confidential information. In this way, certain data may be aggregated without identifying a specific substance, amount, site or corporation. OTS would make the public TSCA CDS available on a publicly accessable national network such as Tymeshare, Lockheed, or SDC. The public TSCA CDS is discussed in more detail in Section 5.2.2.2.

Users would be able to analyze the data on the TSCA CDS utilizing an on-line interactive search and retrieval program. Such a program would provide extensive user prompting in order to allow the novice as well as experienced user to efficiently and effectively retrieve the desired information without elaborate training or knowledge of the system.

A proprietary search and retrieval program would be used exclusively by OTS to access the proprietary TSCA CDS. If OTS implements a proposed plan whereby different units within OTS would be denied access to certain information on the proprietary TSCA CDS, the search and retrieval porgram would maintain the privacy of the data base through verification of a users identity. OTS would establish a set of identification codes based on the various units within OTS and the data on the proprietary TSCA CDS that can be accessed by each unit. These codes would be entered into a Security Data file and accessed by the search and retrieval program when the user presents his/her identity. The proprietary search and retrieval program would maintain an audit trail of all requests for information made against the proprietary TSCA CDS. This audit trail would be analyzed by the data systems contractor in order to monitor the security of the data base. Requests for confidential information that do not originate within OTS, such as Freedom of Information Act and Congressional requests, will be submitted to OTS for review. If the request is valid, OTS will obtain the information from the proprietary TSCA CDS, otherwise, the requester will have to utilize the information on the public TSCA CDS. Security considerations are discussed in more detail in Section 5.2.4.3.

A similar public search and retrieval program would be used to access the public TSCA CDS. This program would be a subset of the proprietary search and retrieval program in that it would not require extensive user identification security procedures and would not maintain an audit trail of user accesses.

5.2.2 Data Base Organization and Structure

The data base proposed by Sigma Data to store the information reported by industry as a result of the Toxic Substances Control Act would consist of the TSCA Chemical Data Systems (CDS) combined with the Reports Management System, both of which are discussed in the Mitre report. Because of the need to protect confidential information, yet provide as much information as possible to the public, two versions of a TSCA CDS would be developed: a proprietary data base and a public data base.

The data base organizations and structure is based on four reporting forms developed by OTS and future reporting requirements as outlined in the Federal Register and EPA strategy documents. The four existing forms relate to TSCA Inventory reporting and include Form A, Form B, Form C, and Form D.

The remainder of this section describes the organization, structure and content of the proposed proprietary and public data bases, and provides an estimate of the storage requirements for the data bases.

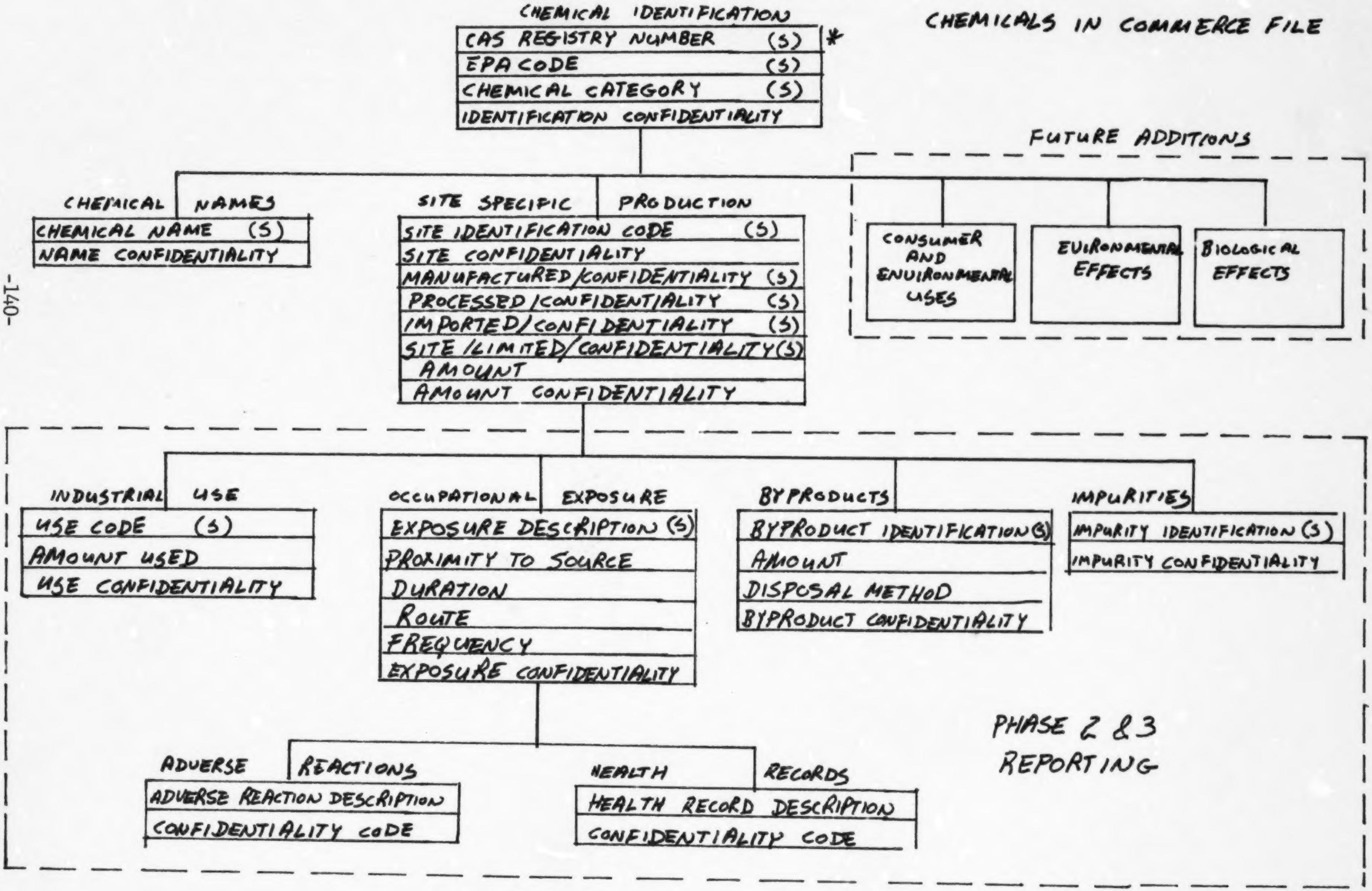
5.2.2.1 Proprietary Data Base

The proposed proprietary TSCA CDS data base would be divided into two files; a Reports Management file and a Chemical In Commerce file. The two files would be directly linked and under the control of a single data base management system (DBMS). Each file would be organized in a hierarchical data structure that will provide an easy means of incorporating future information needs into the overall data base. Figure 5-3 illustrates the hierarchical structure of each file.

The Reports Management file would provide a tracking capability of industrial sites that pertain to TSCA reporting requirements and the individual report submissions made by a particular industrial site. The file would be organized on a site basis and include two information groups or segments; Site Identification and Reports Submitted.

The Site Identification segment would contain information to uniquely identify and locate a particular industrial site. Specific data elements to accomplish this include an identification code, name, address, city, state and zip code. The corporate name and technical contacts are also included in this segment to aid in correspondence and inspection requirements.

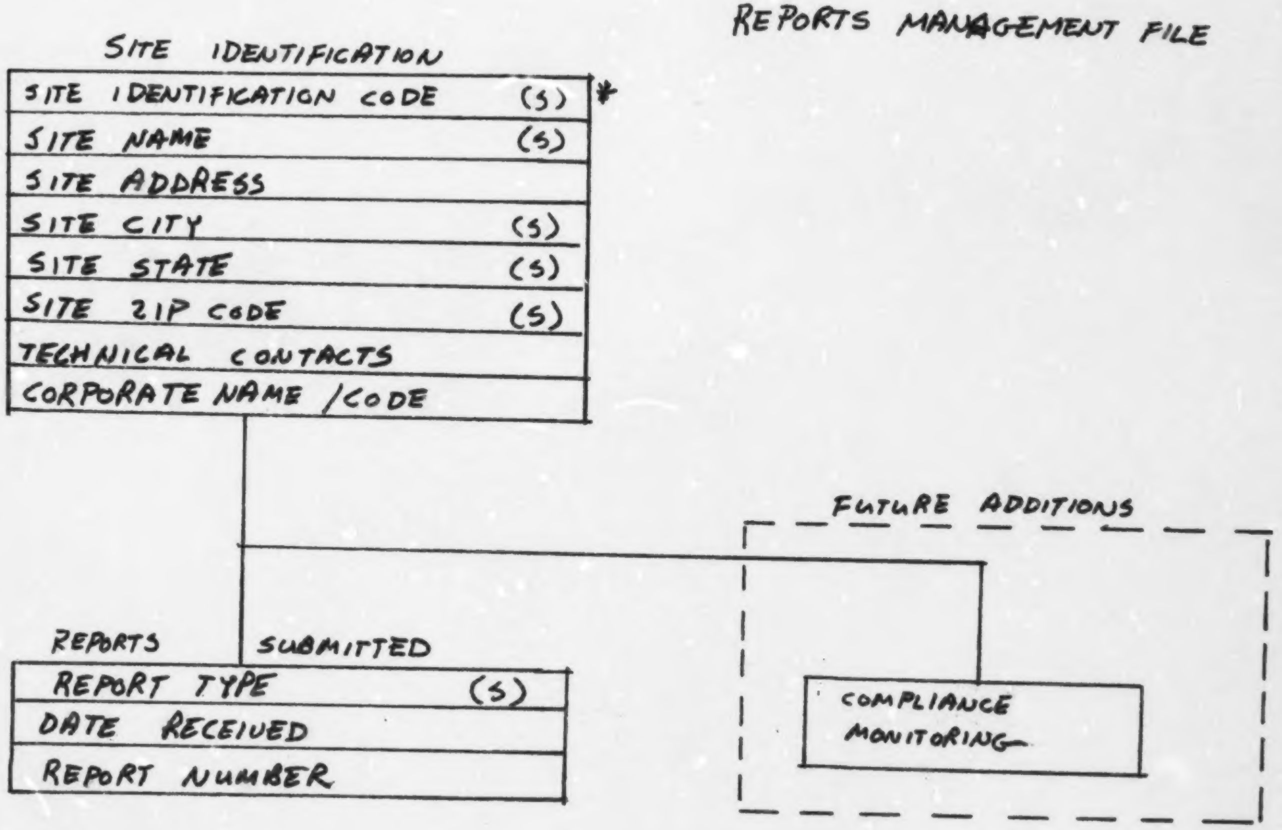
FIGURE 5-3 (1)
PROPRIETARY TSCA CDS DATA BASE STRUCTURE


CHEMICALS IN COMMERCE FILE
CHEMICAL IDENTIFICATION
CAS REGISTRY NUMBER (S) *
EPA CODE (S)
CHEMICAL CATEGORY (S)
IDENTIFICATION CONFIDENTIALITY
FUTURE ADDITIONS
CONSUMER AND ENVIRONMENTAL USES
ENVIRONMENTAL EFFECTS
BIOLOGICAL EFFECTS
CHEMICAL NAMES
CHEMICAL NAME (S)
NAME CONFIDENTIALITY
SITE SPECIFIC PRODUCTION
SITE IDENTIFICATION CODE (S)
SITE CONFIDENTIALITY
MANUFACTURED/CONFIDENTIALITY (S)
PROCESSED/CONFIDENTIALITY (S)
IMPORTED/CONFIDENTIALITY (S)
SITE/LIMITED/CONFIDENTIALITY (S)
AMOUNT
AMOUNT CONFIDENTIALITY
INDUSTRIAL USE
USE CODE (S)
AMOUNT USED
USE CONFIDENTIALITY
OCCUPATIONAL EXPOSURE
EXPOSURE DESCRIPTION (S)
PROXIMITY TO SOURCE
DURATION
ROUTE
FREQUENCY
EXPOSURE CONFIDENTIALITY
BYPRODUCTS
BYPRODUCT IDENTIFICATION (S)
AMOUNT
DISPOSAL METHOD
BYPRODUCT CONFIDENTIALITY
IMPURITIES
IMPURITY IDENTIFICATION (S)
IMPURITY CONFIDENTIALITY
ADVERSE REACTIONS
ADVERSE REACTION DESCRIPTION
CONFIDENTIALITY CODE
HEALTH RECORDS
HEALTH RECORD DESCRIPTION
CONFIDENTIALITY CODE
PHASE 2 & 3
REPORTING

* (S) - DATA ELEMENT IS SEARCHABLE BY USER

FIGURE 5-3 (2)

PROPRIETARY TSCA CDS DATA BASE STRUCTURE


REPORTS MANAGEMENT FILE
SITE IDENTIFICATION
SITE IDENTIFICATION CODE (S) *
SITE NAME (S)
SITE ADDRESS
SITE CITY (S)
SITE STATE (S)
SITE ZIP CODE (S)
TECHNICAL CONTACTS
CORPORATE NAME /CODE
REPORTS SUBMITTED
REPORT TYPE (S)
DATE RECEIVED
REPORT NUMBER
FUTURE ADDITIONS
COMPLIANCE MONITORING

* (S) - DATA ELEMENT SEARCHABLE BY USER

Phase two and three reporting guidelines, specified in the Federal Register indicates that OTS may require industry to submit information on use exposure, byproducts, and impurities for a selected subset of chemicals in commerce. Since the detailed reporting requirements have not yet been established, the following discussion represents a conceptual view of the information segments and their contents that could be incorporated into the database structure.

The Industrial Use segment would contain an entry for each use of a chemical substance at a particular site, the use code would be obtained from the standardized chemical use codes being developed by Auerbach. The amount used and a confidentiality code would also be included in the segment.

An Occupation Exposure segment would provide information about the types of exposure found at a particular site. This segment would contain a description of the exposure, proximity to source, duration, route, frequency, and a confidentiality code. Associated with the Occupational Exposure segment would be information segments for Adverse Reactions and Health Records. Each of these segments would be protected from disclosure by confidentiality codes.

The Byproducts segment would identify each byproduct for a particular chemical substance at a particular site. Byproducts would be identified by name or CAS registry number. Also included in this segment would be the amount of the byproduct, the disposal method and a confidentiality code.

The Impurities segment would identify each impurity associated with a particular chemical at a particular site. A confidentiality code would protect the impurity from unwanted disclosure.

As the information needs of OTS increase, and other TSCA reporting requirements are promulgated by EPA, other information segments will have to be added to the data base structure. Three such future additions are shown in Figure 5-2. Associated with each chemical substance could be a Consumer and Environmental Uses segment, an Environmental Effects segment, and a Biological Effects segment.

5.2.2.2 Public Data Base

OTS has a dual responsibility of making as much information on the proprietary data base available to the public and also protecting the valid confidential claims submitted by industry from unwanted disclosure. The proposed public TSCA CDS data base would be a subset of the proprietary TSCA CDS data base described in the previous section. Figure 5-4 illustrates the hierarchical structure of the public data base. The organization, structure, and content of the public data base is the same as the proprietary data base. The public data base, however, would not contain any confidentiality codes. Since the public user would not be concerned with the tracking reports submitted to OTS, the Reports Management File would be changed to a Site file and the Reports Submitted and Compliance Monitoring information segments would be dropped.

OTS would be responsible for making the public data base available to the general user on a nation-wide timesharing computer network. In order to construct the public TSCA CDS, Sigma Data proposes developing an extract program to process the entire chemicals in Commerce and Reports Management files. The extract program would analyze the various confidentiality codes in a predefined and hierarchical manner. Based on the confidentiality claimed for various data elements, the program would enter the information into the public data base, delete the information from the public data base, or aggregate the information in such a way that specific references or data would be grouped under generic terms. Figure 5-5 illustrates a conceptual processing flow chart of the proposed extract program. The program would process the hierarchical structure from top-to-bottom, left-to-right, and perform the following logical operations:

.If the chemical identity is confidential nothing about the chemical is written to the public data base

.If any chemical names are confidential they are not written to the public data base

.If the existence of a site is confidential nothing about the site is written to the public data base

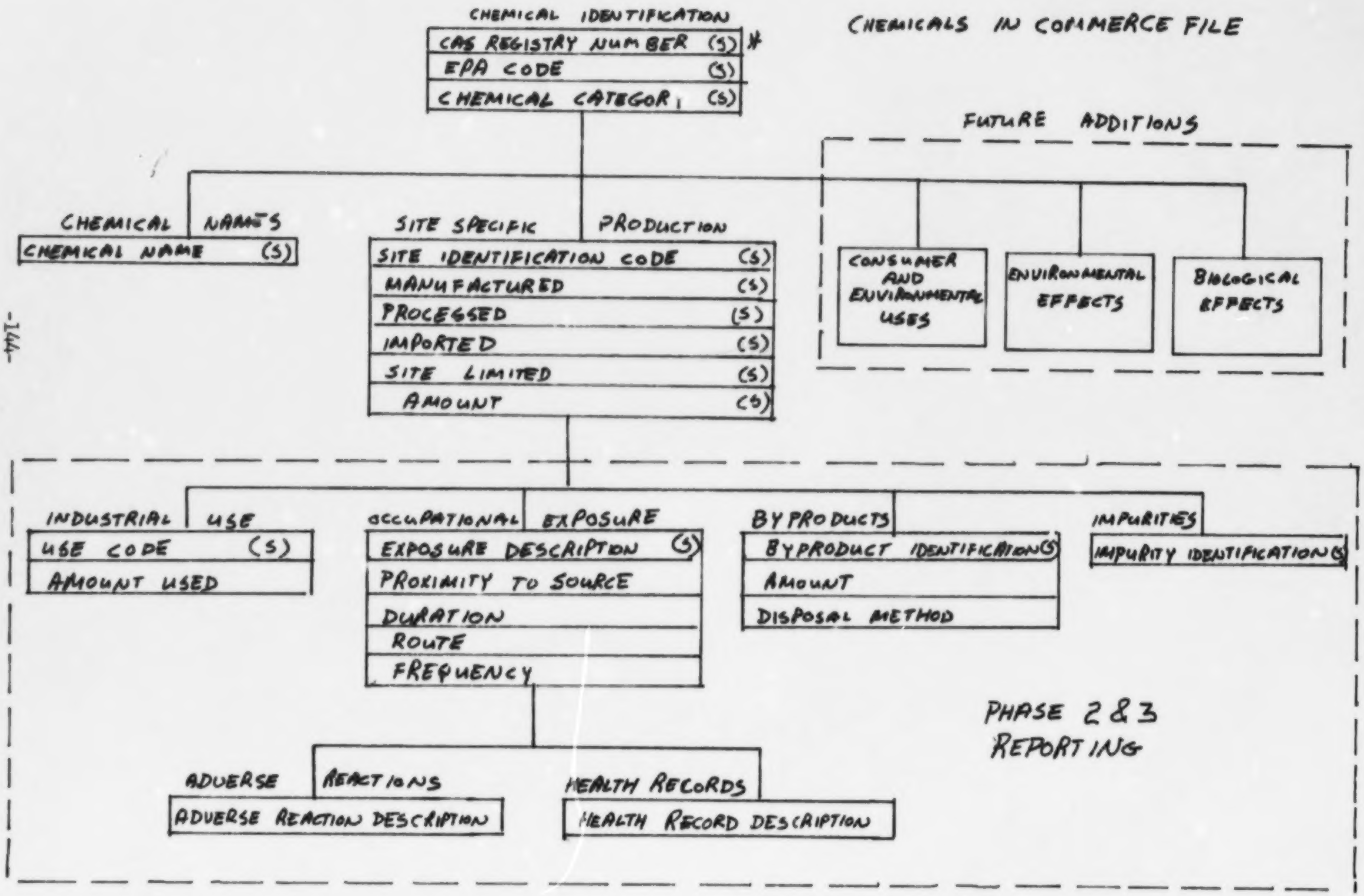
.If specific site location is confidential but data about the site is not, a meaningless generic codes is placed in the site identification code to destroy the acutal site location reference.

.If the manufactured, processed, imported, or site limited indicators are confidential, they are blanked out to destroy any reference.

.If the amount is confidential, a running total of confidential amounts is kept for a

FIGURE 5-4 (1)

PUBLIC TSCA CDS DATA BASE STRUCTURE


CHEMICAL IDENTIFICATION
CAS REGISTRY NUMBER (S) *
EPA CODE (S)
CHEMICAL CATEGORY (S)
CHEMICALS IN COMMERCE FILE
FUTURE ADDITIONS
CONSUMER AND ENVIRONMENTAL USES
ENVIRONMENTAL EFFECTS
BIOLOGICAL EFFECTS
CHEMICAL NAMES
CHEMICAL NAME (S)
SITE SPECIFIC PRODUCTION
SITE IDENTIFICATION CODE (S)
MANUFACTURED (S)
PROCESSED (S)
IMPORTED (S)
SITE LIMITED (S)
AMOUNT (S)
INDUSTRIAL USE
USE CODE (S)
AMOUNT USED
OCCUPATIONAL EXPOSURE
EXPOSURE DESCRIPTION (S)
PROXIMITY TO SOURCE
DURATION
ROUTE
FREQUENCY
BYPRODUCTS
BYPRODUCT IDENTIFICATION (S)
AMOUNT
DISPOSAL METHOD
IMPURITIES
IMPURITY IDENTIFICATION (S)
ADVERSE REACTIONS
ADVERSE REACTION DESCRIPTION
HEALTH RECORDS
HEALTH RECORD DESCRIPTION
PHASE 2 & 3 REPORTING

* (S) - DATA ELEMENT SEARCHABLE BY USER

FIGURE 5-4 (2)

PUBLIC TSCA CDS DATA BASE STRUCTURE

SITE FILE

SITE IDENTIFICATION

SITE IDENTIFICATION CODE	(S) *
SITE NAME	(S)
SITE ADDRESS	
SITE CITY	(S)
SITE STATE	(S)
SITE ZIP CODE	(S)
COPORATE NAME	

* (S) - DATA ELEMENT IS SEARCHABLE BY USER

FIGURE 5-5 (1)

PUBLIC TSCA CDS EXTRACT FLOW CHART

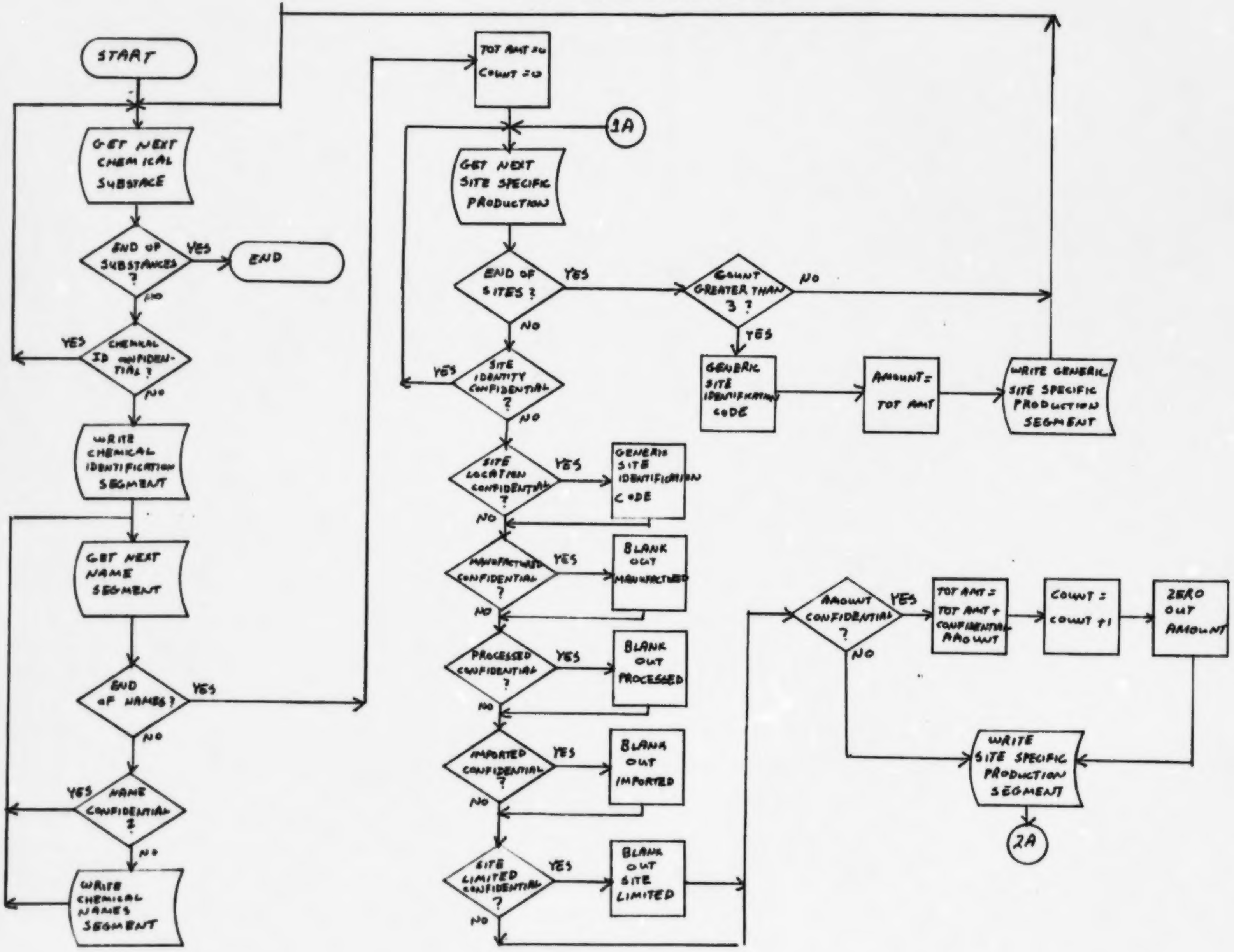
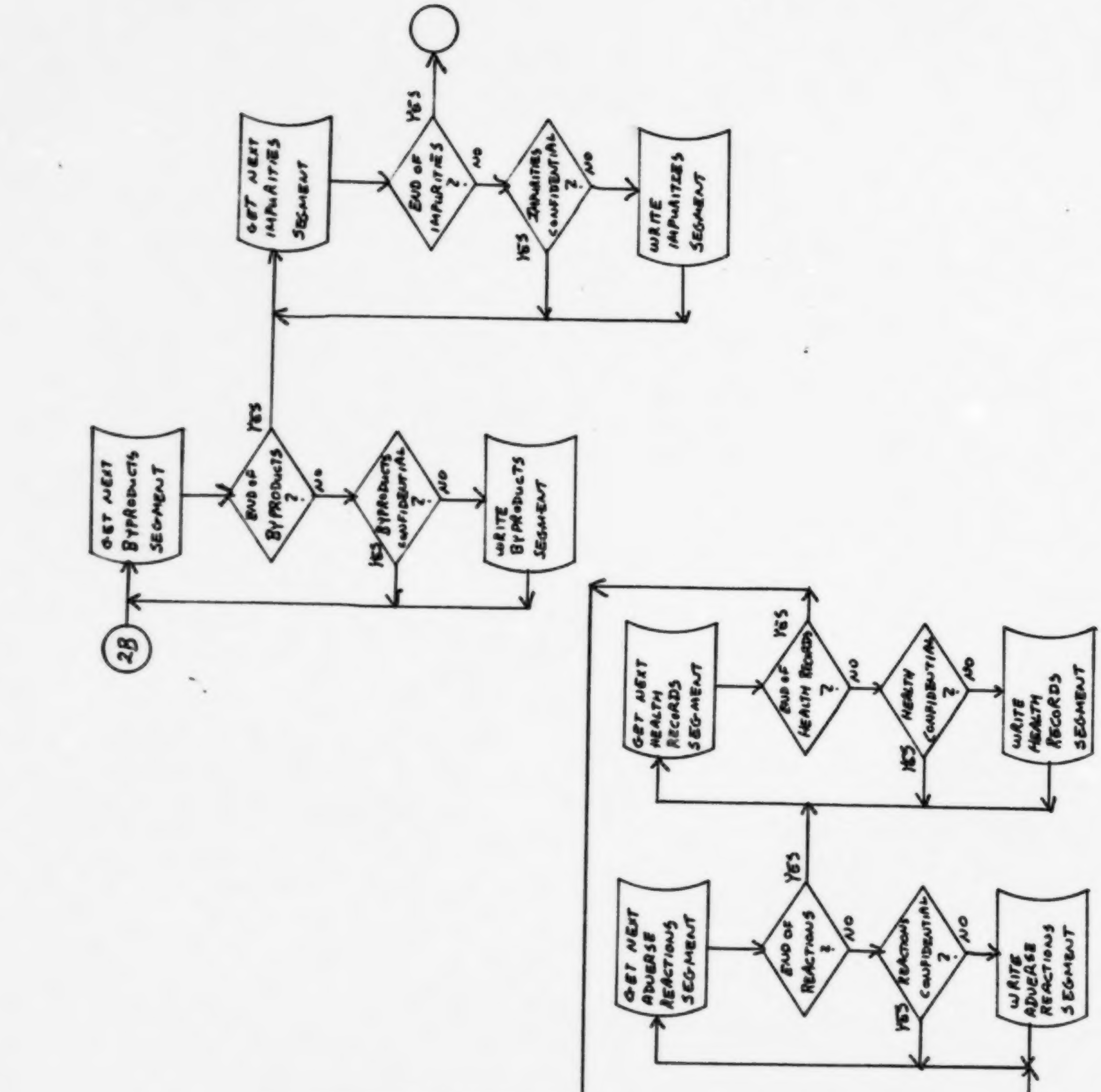

START
GET NEXT CHEMICAL SUBSTACE
END OF SUBSTANCES ?
YES
END
NO
CHEMICAL ID CONFIDENTIAL ?
YES
NO
WRITE CHEMICAL IDENTIFICATION SEGMENT
GET NEXT NAME SEGMENT
END OF NAMES ?
YES
NO
NAME CONFIDENTIAL ?
YES
NO
WRITE CHEMICAL NAMES SEGMENT
TOT AMT = 0
COUNT = 0
1A
GET NEXT SITE SPECIFIC PRODUCTION
END OF SITES ?
YES
NO
COUNT GREATER THAN 3 ?
NO
YES
GENERIC SITE IDENTIFICATION CODE
AMOUNT = TOT AMT
WRITE GENERIC SITE SPECIFIC PRODUCTION SEGMENT
SITE IDENTITY CONFIDENTIAL ?
YES
NO
SITE LOCATION CONFIDENTIAL ?
YES
NO
GENERIC SITE IDENTIFICATION CODE
MANUFACTURED CONFIDENTIAL ?
YES
NO
BLANK OUT MANUFACTURED
PROCESSED CONFIDENTIAL ?
YES
NO
BLANK OUT PROCESSED
IMPORTED CONFIDENTIAL ?
YES
NO
BLANK OUT IMPORTED
SITE LIMITED CONFIDENTIAL ?
YES
NO
BLANK OUT SITE LIMITED
AMOUNT CONFIDENTIAL ?
YES
NO
TOT AMT = TOT AMT + CONFIDENTIAL AMOUNT
COUNT = COUNT + 1
ZERO OUT AMOUNT
WRITE SITE SPECIFIC PRODUCTION SEGMENT
2A

FIGURE 5-5 (2)



2A
GET NEXT INDUSTRIAL USE SEGMENT
END OF USES ?
YES
NO
USE CONFIDENTIAL ?
YES
NO
WRITE INDUSTRIAL USE SEGMENT
GET NEXT OCCUPATIONAL EXPOSURE SEGMENT
END OF EXPOSURES ?
YES
2B
NO
EXPOSURE CONFIDENTIAL ?
YES
NO
WRITE OCCUPATIONAL EXPOSURE SEGMENT
GET NEXT ADVERSE REACTIONS SEGMENT
END OF REACTIONS ?
YES
NO
REACTIONS CONFIDENTIAL ?
YES
NO
WRITE ADVERSE REACTIONS SEGMENT
GET NEXT HEALTH RECORDS SEGMENT
END OF HEALTH RECORDS ?
YES
NO
HEALTH CONFIDENTIAL ?
YES
NO
WRITE HEALTH RECORDS SEGMENT
2B
GET NEXT BYPRODUCTS SEGMENT
END OF BYPRODUCTS ?
YES
NO
BYPRODUCTS CONFIDENTIAL ?
YES
NO
WRITE BYPRODUCTS SEGMENT
GET NEXT IMPURITIES SEGMENT
END OF IMPURITIES ?
YES
NO
IMPURITIES CONFIDENTIAL ?
YES
NO
WRITE IMPURITIES SEGMENT

particular chemical and a count is kept of how many sites claim a confidential amount for a particular chemical. If more than three sites for a particular chemical claim a confidential amount, a Site Specific Production segment is written to the public data base with a generic site identification code and the running total of the confidential amount. The public user will not be able to connect the amount with any particular site.

- If any industrial uses are confidential they are not written to the public data base.
- If an occupational exposure is confidential, the Occupational Exposure segment, and Adverse Reactions and Health Records segments associated with it are not written to the public data base.
- If any adverse reactions or health records are confidential they are not written to the public data base.
- If any byproducts or impurities are confidential, they are not written to the public data base.

The Site file of the Public TSCA CDS would be constructed by extracting only the site identification segment of the Reports Management file of the proprietary TSCA CDS.

5.2.2.3 Storage Estimates

This section presents estimates of the storage requirements over a five year period for the proprietary and public data bases of the TSCA Chemical Data System. These estimates are intended as guidelines to be used in selecting computers and storage devices for the actual implementation of the TSCA CDS.

Table 5-1 indicates the size of each data element in the proprietary and public data bases. The total for each information segment in the hierarchical structure of the Reports Management and Chemicals in Commerce files is also given. The size of each data element does not take into consideration any compaction strategies that might be employed in the actual implementation.

TABLE 5-1 (1)

TSCA CDS

DATA ELEMENT SIZE

REPORTS MANAGEMENT FILE

Data Element	Proprietary	Public
Site Identification		
Site Identification Code	9	9
Site Name	30	30
Site Address	60	60
Site City	30	30
Site State	2	2
Site Zip Code	5	5
Technical contacts	30	
Corporate Name	30	
Total	196	136
Reports Submitted		
Report Type	2	
Date Received	6	
Report Number	7	
Total	15	

CHEMICALS IN COMMERCE FILE

Data Element	Proprietary	Public
Chemical Identification		
CAS Registry Number	9	9
EPA Number	6	6
Chemical Class	4	4
Identification Confidentiality	1	
Total	20	19

TABLE 5-1 (2)

Data Element	Proprietary	Public
Chemical Names		
Chemical Name	60	60
Name Confidentiality	1	
Total	61	60
Site Specific Production		
Site Identification Code	9	9
Site Confidentiality	1	
Manufacture Indicator/Confidentiality	1	1
Process Indicator/Confidentiality	1	1
Import Indicator/Confidentiality	1	1
Site Limited Indicator/Confidentiality	1	1
Amount	9	9
Amount Confidentiality	1	
Total	24	22
Industrial Use		
Use Code	8	8
Amount Used	9	9
Use Confidentiality	1	
Total	18	17
Occupational Exposure		
Exposure Description	8	8
Exposure Confidentiality	1	
Proximity to Source	5	5
Duration	6	6
Route	6	6
Frequency	6	6
Total	32	31

TABLE 5-1 (3)

Data Element	Proprietary	Public
Adverse Reactions		
Reaction Description	30	30
Confidentiality Code	1	
Total	31	30
Health Records		
Record Description	30	30
Confidentiality Code	1	
Total	31	30
Byproducts		
Byproduct Identification	60	60
Byproduct Confidentiality	1	
Amount	9	9
Disposal Method	10	10
Total	80	79
Impurities		
Impurity Identification	60	60
Impurity Confidentiality	1	
Total	61	60

Table 5-2 presents the storage requirements of the proprietary TSCA CSA over a five year period. Each information segment is identified along with its total length. The volume indicates the cumulative number of iterations for a particular information segment within a particular year. The total number of characters for an information segment is the product of the volume times the length. The total number of characters required to store the proprietary TSCA CDS data for a particular year is shown on the bottom line.

Year 1 assumes that phase one of the TSCA reporting requirements will be in effect as outlined in the August 2, 1977 Federal Register. In this phase, OTS will collect site specific production information for an estimated 50,000 chemical substances included in the TSCA inventory of chemicals in Commerce. OTS estimates that 20,000 sites would be required to submit production information. This is reflected in the volume for the Site Identification segment and the production portion of the Reports Submitted section. It is estimated that each chemical on the inventory would have an average of three names associated with it and that an average of seven sites would submit site specific production information for each chemical substance. During year 1, premarket notification will be initiated, and it is estimated that OTS would review about 1,000 new chemical substances. Based on the volumes for the various information segments described above, the total number of characters for year 1 would be about 23 million.

It is assumed that the number of sites would increase at the rate of 5 percent a year. This is reflected in the volume increases for the Site Identification, Site Specific Production, and the production portion of the Reports Submitted sections. Premarket notification would remain the same each year.

Year 2 assumes that the second phase of the TSCA reporting requirements would be in effect. In this phase, OTS would require additional information for an estimated 2,000 substances selected from the TSCA inventory of chemicals in Commerce. The information would include use, byproducts, impurities and exposure data. It is estimated that 3,000 sites would be required to submit reports for this additional information. This is reflected in the Reports Submitted segment. Since an average of seven sites would report information for each of the 2,000 selected chemicals, a total of 14,000 Site Specific Production segments would reference additional information segments in the hierarchical structure. For each of the 14,000

TABLE 5-2

TSCA CHEMICAL DATA SYSTEM STORAGE ESTIMATES

Information Segment	Length	Year 1		Year 2		Year 3		Year 4		Year 5	
		Volume	Charact.	Volume	Charact.	Volume	Charact.	Volume	Charact	Volume	Charact.
Reports Management											
Site Identification	196	20,000	3,920,000	21,000	4,116,000	22,050	4,321,800	23,150	4,537,400	24,340	4,770,640
Reports Submitted											
·Production	15	20,000	300,000	21,000	315,000	22,050	330,750	23,150	347,250	24,340	365,100
·Premarket Notif.	15	1,000	15,000	2,000	30,000	3,000	45,000	4,000	60,000	5,000	75,000
Use	15			3,000	45,000	15,000	225,000	15,750	236,250	16,530	247,950
·Byproducts and Impurities	15			3,000	45,000	15,000	225,000	15,750	236,250	16,530	247,950
·Exposure	15			3,000	45,000	15,000	225,750	15,750	236,250	16,530	247,950
Chemicals in Commerce											
Chemical Ident.	20	50,000	1,000,000	51,000	1,020,000	52,000	1,040,000	53,000	1,060,000	54,000	1,080,000
Chemical Names	61	150,000	9,150,000	153,000	9,333,000	156,000	9,516,000	159,000	9,699,000	162,000	9,882,000
Site Specific Prod.	24	350,000	8,400,000	367,500	8,820,000	385,880	9,261,120	405,170	9,724,080	425,430	10,210,320
Industrial Use	18			56,000	1,008,000	140,000	2,520,000	147,000	2,646,000	154,350	2,778,300
Occupational Exposure	32			28,000	896,000	70,000	2,240,000	73,500	2,352,000	77,180	2,469,760
Adverse Reactions	31			56,000	1,736,000	140,000	4,340,000	147,000	4,557,000	154,350	4,784,850
Health Records	31			140,000	4,340,000	370,000	11,470,000	388,500	12,043,500	407,930	12,645,830
Byproducts	80			42,000	3,360,000	105,000	8,400,000	110,250	8,820,000	115,760	9,260,800
Impurities	61			42,000	2,562,000	105,000	6,405,000	110,250	6,725,250	115,760	7,061,360
TOTAL			22,785,000		37,671,000		60,564,670		63,280,230		66,127,810

sites there would be an estimated 5 Industrial Use segments, 2 Occupational Exposure segments, 3 Byproducts segments, and 3 Impurities segments. For each Occupational Exposure segment there would be 5 Adverse Reactions segments and 4 Health Records segments. Using the above volumes, an estimated 38 million characters would be required in year 2.

Year 3 assumes that the third phase of the TSCA reporting requirements would be in effect. In this phase, OTS would require the same type of information, but the number of chemicals selected would increase to 5,000. It is estimated that a total of 15,000 would increase to 5,000. It is estimated that a total of 15,000 would have to submit reports for use, byproducts, impurities and exposure data. The number of Site Specific Production segments affected would be approximately 35,000. The number of Industrial Use, Occupational Exposure, Byproducts, Impurities, Adverse Reactions, and Health Records segments associated with each site would be the same as described in Year 2. Using the above volumes, an estimated 61 million characters would be required.

OTS has not, as yet, formulated any new TSCA reporting requirements past year 3. Therefore, the volume of data should remain constant except for the assumed annual 5 percent increase in commercial sites. This is reflected in the total storage required for years 4 and 5 which is 63 million and 66 million characters respectively.

The storage estimates in Table 5-2 are for data only. The extra storage required to hold indexing and pointer information for the various data elements that are searchable was not included. This extra storage depends upon the data base management system (DBMS) selected to actually implement the TSCA CDS. Most DBMS's require 50 to 100 percent of the total storage devoted to data for indexing information.

The estimate also assumes that the sites reporting information would be restricted to manufactures and remains relatively constant at around 20,000. If OTS decides to require extensive information from processors, the number of commercial sites reporting information will increase significantly. The majority of information in the data base contains data about a particular site. An increase of 100 percent in the number of sites required to report TSCA information would cause about an 85 percent increase in the storage requirements for the TSCA CDS.

Table 5-2 only gives storage estimates for the proprietary data base of the TSCA CDS. The public data base of the TSCA CDS would have the same organization and structure as the proprietary data base, but would not contain tracking data about reports submitted, corporate names, technical contacts, or confidentiality codes. The public data base would not contain any information about chemicals declared confidential, which, in the Mitre report, were estimated at 1 in a 1,000. Information concerning industrial use, occupational exposure, byproducts, impurities and site location that was declared confidential, but for a non-confidential chemical, will not be included in the public data base. The entire amount of confidential information should not reduce the size of the public data base by more than 15 percent.

5.2.3 Search and Retrieval Capabilities

Search and retrieval of TSCA CDS information would be accomplished through a user-prompted dialogue between a user entering information on a terminal and a computer program accessing the TSCA CDS data base. Since the proprietary and public TSCA CDS data base have basically the same organization and structure, the same software could be used to search both data bases.

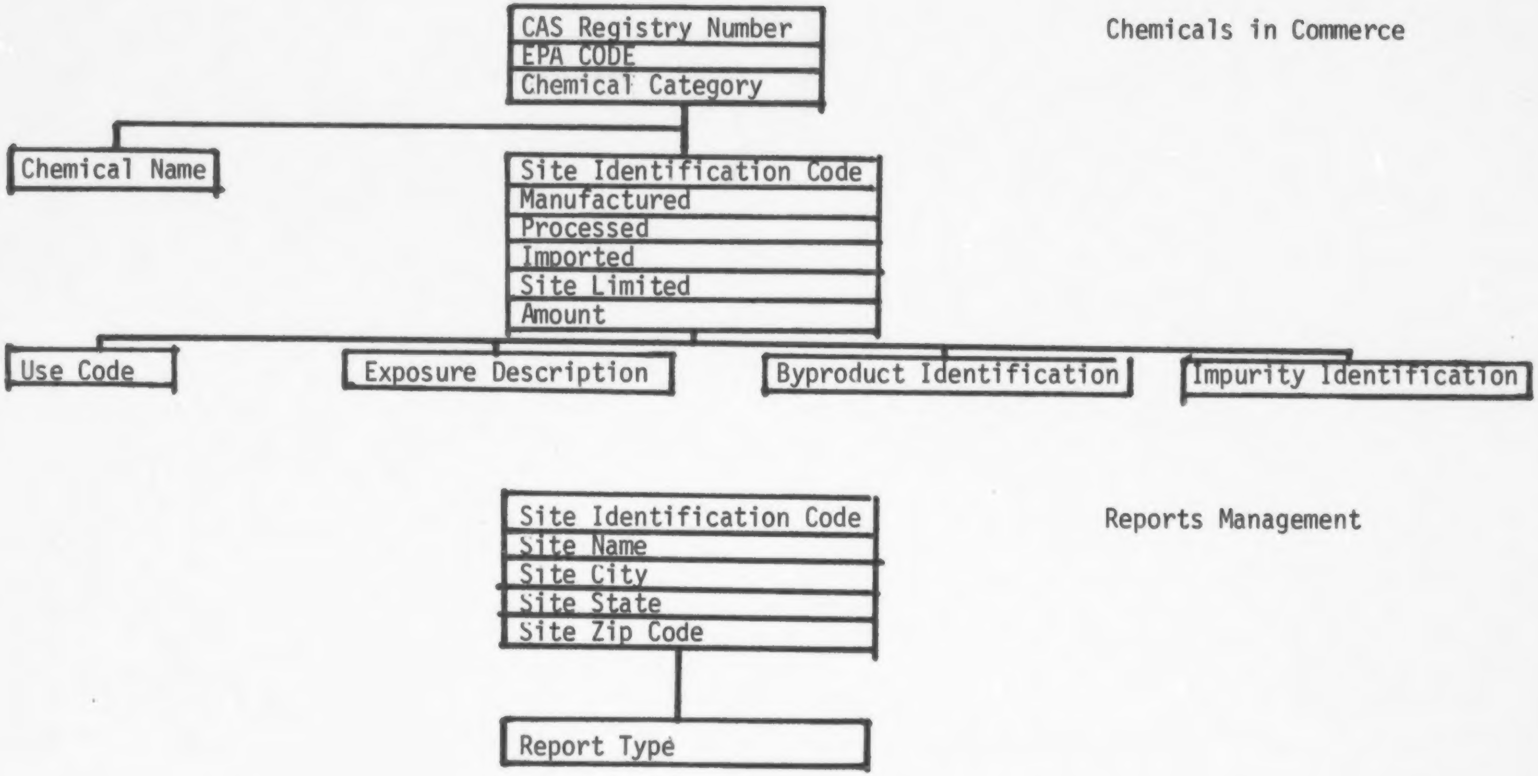
Since the TSCA CDS would be accessed by many inexperienced users, the search and retrieval program should provide a series of "help" commands that would explain how the system operates, describe the various data elements in the data base, and describe the structure of the various commands used to access the system.

The hierarchical structure of the TSCA CDS provides the user with a wide range of data elements that could be indexed and used in a variety of search and retrieval query specifications. Figure 5-6 is a subset of Figure 5-3 and illustrates the minimum data elements in the hierarchy that should be indexed to provide direct search and retrieval of data base information. These data elements would be used as the main search criteria when developing a query specification. Other data elements that are not indexed could be used to provide a secondary search limitation. For instance, from Figure 5-3, use code is indexed while amount used is not. The search and retrieval capabilities of the TSCA CDS should allow the user to specify the following query:

- what commercial sites manufacture chemicals for use code 3562 in amounts greater than 10,000 lbs.

FIGURE 5-6

TSCA CDS INDEXED DATA ELEMENTS


CAS Registry Number
EPA CODE
Chemical Category
Chemicals in Commerce
Chemical Name
Site Identification Code
Manufactured
Processed
Imported
Site Limited
Amount
Use Code
Exposure Description
Byproduct Identification
Impurity Identification
Site Identification Code
Site Name
Site City
Site State
Site Zip Code
Reports Management
Report Type

The search criteria specified by the user would be matched against various information segments in the TSCA CDS data base. The results of a query would be stored in a "save" file and the number of matches will be returned to the user. A number of "save" files, each with its own unique identifier, would be available to the user to provide multiple searches within the same session. Upon receiving the results of a search the user would have the option to:

- refine the search query
- display the results of the search
- initiate a new query

The search query would be refined by combining the contents of various "save" files using logical connectives (AND, OR, NOT), or by specifying further restrictions on the "save" file using conditional data element statements. The user would be able to display the results of a search in several different ways, The options include printing specified data elements, information segments, or complete data base records.

Based on the above discussion, the search and retrieval program should be capable of providing answers to the following types of questions:

- What sites manufacture a particular chemical?
- What chemicals are processed at a particular site?
- What are the chemical names of the substance with CAS Registry Number 361-42-8?
- For a particular chemical, what sites manufacture more than 10,000 lbs.?
- What is the occupational exposure and associated adverse reactions for a particular chemical?
- What sites have submitted Form C? (Proprietary TSCA CDS only)
- What commercial sites in Delaware use more than 3,000 lbs. of a particular chemical under use code 63991?

- What chemical identities are considered confidential? (proprietary TSCA CDS only)

- What byproducts are associated with a particular chemical?

5.2.4 Implementation Alternatives

The implementation plan adopted for the TSCA Chemical Data System will have a major impact on the future information collection, storage, and dissemination capabilities of OTS. It is therefore important that the right decisions be made at the beginning of the system development process. In the case of the TSCA CDS, this is difficult because many of the detailed specifications required to develop a comprehensive system are not well defined or not defined at all. Therefore, many unanticipated OTS needs will have to be incorporated into the system at a later date. The TSCA CDS must be responsive to the immediate needs of OTS and yet be flexible enough to accomodate future requirements with a minimum amount of reorganization and redesign.

Based on the operational, data base, storage, and input/output requirements of the TSCA Chemical Data Base System discussed in Sections 5.2.1, 5.2.2, and 5.2.3, Sigma Data determined that the system could be implemented on most medium to large scale timeshared computers as well as fully equipped minicomputers. Therefore, two alternative plans are presented for implementing the TSCA CDS which include utilizing a dedicated minicomputer or a large-scale timesharing computer. Since security plays an important role in the TSCA CDS, it is discussed separately. Finally, the advantages and disadvantages of each plan are summarized to provide a comparison between the two alternatives.

5.2.4.1 Plan 1 - Dedicated Minicomputer

The first alternative for implementing the TSCA CDS involves utilizing a dedicated minicomputer to process the information submitted to OTS. The processing powers of minicomputers have been increasing at a tremendous rate in the past few years. Most minicomputers can be equipped with sophisticated virtual storage teleprocessing operating systems that could adequately service as many as 32 on-line terminals. Auxilliary storage capacity has expanded to where 500 million characters could easily be stored on a minicomputer system. Core storage capacity has increased correspondingly to where 250 K bytes and over is not uncommon. These characteristics should be adequate for TSCA CDS processing requirements. The trend is toward even greater processing power, to where a fully loaded minicomputer resembles a medium-scale mainframe computer.

The minicomputer compares favorably to the mainframe in price and installation requirements. The purchase price and operational costs are less for a minicomputer than for a mainframe. Minicomputers do not require elaborate power supplies and sophisticated air conditioning systems.

The minicomputer, however, does have certain limitations compared to a mainframe. Supporting software and utilities that are provided with most mainframes are lacking or not fully developed for certain minicomputers. Data base management systems, that are readily available on most mainframe computers, are not as readily available on minicomputers, and most provide only a subset of the full capabilities available on a mainframe. Some minicomputer programming languages only implement a subset of the full language capabilities found on the larger mainframe computers.

The major advantage of a minicomputer is that the user has greater control over the computer and the system can be dedicated to a particular user need. The minicomputer system could be configured to meet the specific requirements of the user and can easily be upgraded in the future to meet additional user needs. Since the minicomputer could be dedicated to the TSCA CDS processing requirements, certain privacy and security considerations would be inherrent in the stytem. The operating system would only process a single application, the TSCA CDS. The only on-line access to the system would be by OTS personnel. A more detailed discussion of security considerations is presented in Section 5.2.4.3.

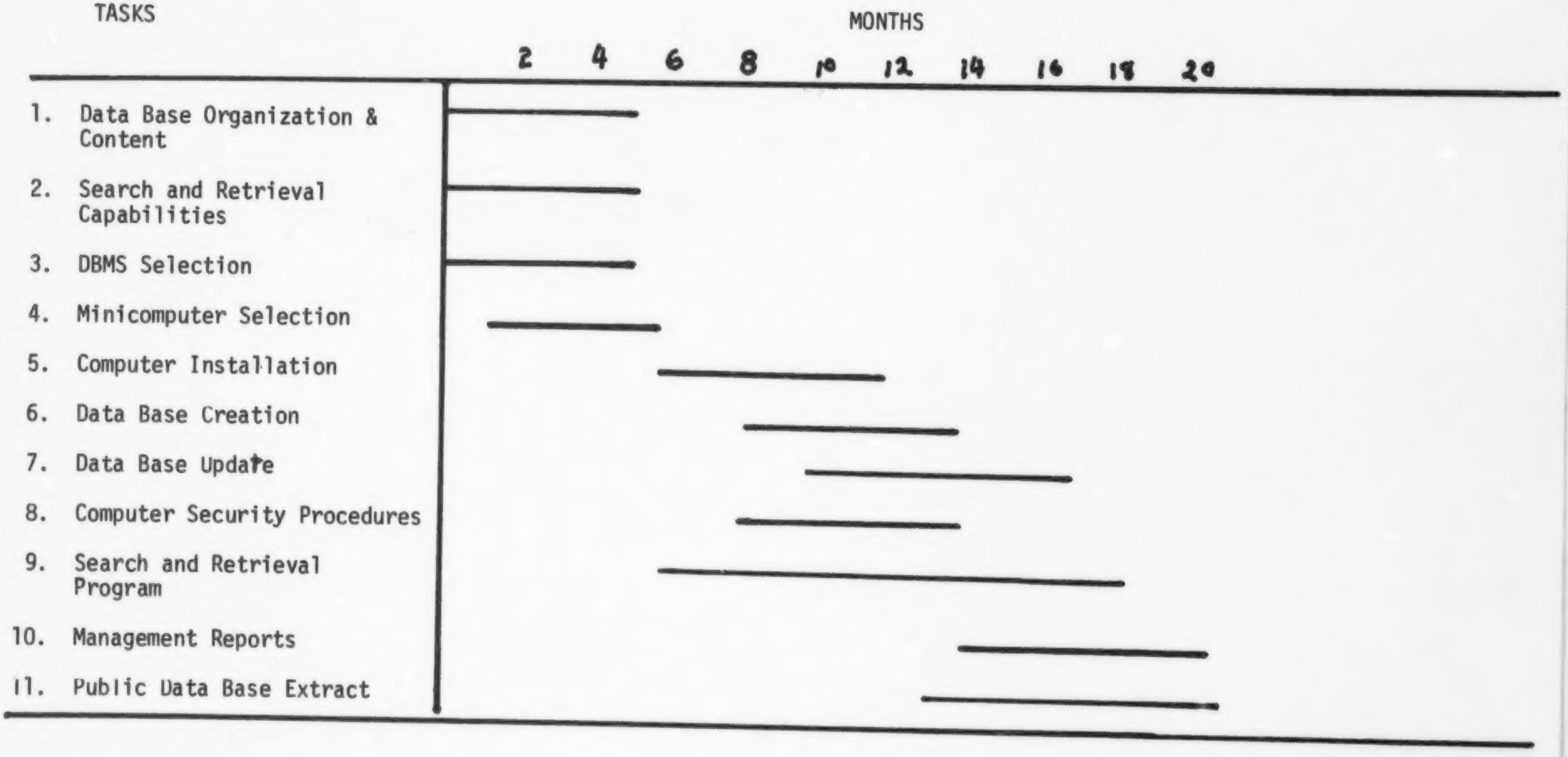
Table 5-3 lists the tasks required to implement the TSCA CDS on a dedicated minicomputer and a time phased schedule for completion. The individual tasks are described below.

Task 1 - Data Base Organization and Content

Based on the general data base organization and content discussed in Section 5.2.2., a more detailed organization and content to be used in the actual implementation would have to be defined. Phase 2 and 3 reporting requirements, which are as yet undefined, would be developed by OTS and incorporated into the data base hierarchy. The physical characteristics of each data element in the data base would also be defined. Future OTS needs that are not yet well defined should be addressed and consideration given as to how these needs could be incorporated into the overall data base design.

TABLE 5-3

TSCA CDS PLAN I IMPLEMENTATION TASKS



Task 2 - Search and Retrieval Capabilities

Based on the general search and retrieval capabilities discussed in Section 5.2.3 and the organization and content developed in Task 1, detailed search and retrieval requirements would be defined. In this task, the data elements that could be searched directly, the various types of search strategies that may be employed, and the format for displaying search results would be specified. The search and retrieval capabilities of the EPA regions would also be defined in this task. These capabilities would inlcude either on-line interactive or telephone requests to OTS for information on the proprietary TSCA CDS.

Task 3 - DBMS Selection

A survey would be conducted of data base management systems available for minicomputers that met the requirements defined in Tasks 1 and 2. If the survey indicates that the requirements could not be met by any DBMS, the data base organization and search and retrieval requirements may be modified in order to accommodate a particular DBMS. However, the needs of OTS should not be bypassed for the sake of implementing a particular data base management system. If a DBMS cannot be found, the data base organization and maintenance would have to be implemented using custom written software.

Task 4 - Minicomputer Selection

A minicomputer would be selected based on the data base organization, storage requirements, search and retrieval capabilities, operational requirements, and communications requirements discussed in Sections 5.2.1, 5.2.2, and 5.2.3 and Tasks 1, 2 and 3 above. Both short and long-term OTS requirements would be considered in the minicomputer selection.

Within EPA there are several operational minicomputer systems. EPA has developed a minicomputer policy and has published guidelines concerning minicomputer feasibility studies and implementation procedures. A standard EPA minicomputer configuration has been approved which consists of a DEC PDP 11/70 with a real-time operating system, 256K words of core storage, COBOL compiler, 16 on-line terminals, and two 88M byte disk packs. This configuration could be upgraded based on a users needs, but should be sufficient for the short-term needs of OTS. Because of this EPA approved minicomputer procurement policy, it is recommended that prime consideration be given to developing the proprietary TSCA CDS according to the standard configuration or an upgraded version of the standard configuration. A detailed minicomputer feasibility study should be conducted according to EPA specified guidelines. If the study reveals that the DEC PDP 11/70 could not meet the requirements of the proprietary TSCA CDS, a survey of other minicomputers that could meet the requirements should be conducted. Some other minicomputers that might be analyzed include the Data General Eclypse, the Hewlett Packard 300cx, or the Interdata 8/32.[1] These machines would provide the processing power required by the TSCA CDS.

The results of this task may depend on the results of Task 4, the DBMS selection. If a data base management system is found that satisfies the requirements of the TSCA CDS, it may be restricted to one or two computers that may not include the EPA standard configuration or another machine selected. In this case, a decision would have to be made to modify the TSCA CDS requirements or develop the TSCA CDS without using a data base management package.

Task 5 - Computer Installation

Once the minicomputer has been selected, it must be procurred and installed. A decision would have to be made whether EPA or the data system contractor for OTS would provide the computer. If EPA provides the computer, a purchase versus lease decision would be made according to EPA guidelines. The specific computer site would also be selected. Since privacy and security could be enhanced through direct connection between on-line terminals and the minicomputer, prime consideration should be given to installing the minicomputer in close proximity to the Office of Toxic Substances (500 to 1,000 ft.). A decision would also be made regarding systems operations, which includes whether EPA or data system contractor personnel would be used to run the system. A typical minicomputer would require an operator, a supervisor, and a system/application programmer. If the EPA standard configuration is selected, installation would take around six months. If another minicomputer is selected, implementation would probably take a few months longer.

---

[1] Auerbach Guide to Minicomputers, 1976, pp. 96, 252, 275.

Task 6 - Data Base Creation

Chemical Abstracts Service (CAS) will create the initial Inventory of Toxic Substances along with the site specific production reporting information collected under Phase 1 of the TSCA reporting requirements. CAS will provide the data systems contractor responsible for TSCA CDS with tapes of the data collected. A special program would be developed to construct the initial version of the Chemicals in Commerce, and Reports Management files of the TSCA CDS.

Task 7 - Data Base Update

Update procedures would be designed and developed to apply the information collected by OTS to the TSCA CDS. The information would include information on premarket notification, site specific production, industrial user, occupational exposure, byproducts and impurities. Procedures would also be developed to maintain the Site Identification information segment of the Reports Management file. If a data base management system is selected, any on-line update capabilities supplied as part of the package should be considered. If these facilities are not available a batch type updating procedure would be sufficient based on the volume and response time required for data.

Task 8 - Computer Security Procedures

Security measures would be defined and developed to protect the confidential information from unwanted disclosure or access. Section 5.2.4.3, discusses in more detail the security considerations involved in the two implementation alternatives.

Task 9 - Search and Retrieval Program

A user-prompted search and retrieval program would be designed and developed to satisfy the requirements developed in Task 2.

Task 10 - Management Reports

Various statistical and status reports that would be used by management to monitor and control the data base would be designed and developed.

Task 11 - Public Data Base Extract

Based on the data base organization developed for the public TSCA CDS and the confidentiality designation for various data elements defined in Task 1, for the proprietary TSCA CDS, a computer program would be developed to construct the public TSCA CDS. The program would be based on the procedures proposed in Section 5.2.2.2.

As table 5-3 indicated an initial version (Phase I TASCA reporting) of the proprietary TSCA CDS could be available on a minicomputer within 14 months. A fully operational version including Phase II reporting data, full search and retrieval capabilities, management reporting data, and a public version of the data base could be available within 20 months. During the six month period required for computer installation computer program development for data base creation, updating, and search and retrieval could be conducted using a timesharing service bureau having the same make minicomputer, programming language software, and data base management system (if selected).

It is estimated that the 11 implementation tasks would require 130 man months of effort and cost around $525,000. The level of effort would be divided between project managers, systems analysts, and programmers, and does not include the personnel required to operate the minicomputer system. Operational support would include, at a maximum, an operator, supervisor, and systems programmer.

The purchase price for the EPS standard DEC PDP 11/70 minicomputer including installation, is estimated at $250,000. If a minicomputer other than the PDP 11/70 were selected, the cost would be relatively close to the $250,000 figure.

5.2.4.2 Plan II - Large-Scale Time-Shared Computer

The second alternative for implementing the TSCA CDS involves utilizing a large-scale time-shared computer to process the information reported under TSCA. EPA currently utilizes several large scale timesharing computers to implement many of their applications. These computers provide the user with a wide variety of software, such as data base management systems text editors, libraries, and file management utilities. This abundance of software can reduce development time and decrease development costs.

Since the computer resources and costs are shared by many users better utilization of the system is realized. A system such as the TSCA CDS would be small at first and require a minimum amount of processing time and storage. Under this plan, the TSCA CDS would only be billed for the actual resources used, yet still having access to the full computing power of a large machine. The TSCA CDS could grow without having to periodically upgrade a dedicated computer system.

A time-share computer system does have its drawbacks as far as security and privacy are concerned. In such a system it is much more difficult to monitor and control users that attempt to access confidential information. A more detailed discussion of the security involved in such a system is presented in Section 5.2.4.3.

Table 5-4 lists the tasks required to implement the TSCA CDS on a large scale timesharing computer. A time phased schedule for implementation is also included. Each task is discussed below:

Task 1 - Data Base Organization and Content

This task is identical to Task 1 described in Section 5.2.4.1.

Task 2 - Search and Retrieval Capabilities

This task is identical to Task 2 described in Section 5.2.4.1.

Task 3 - Computer/Installation Selection

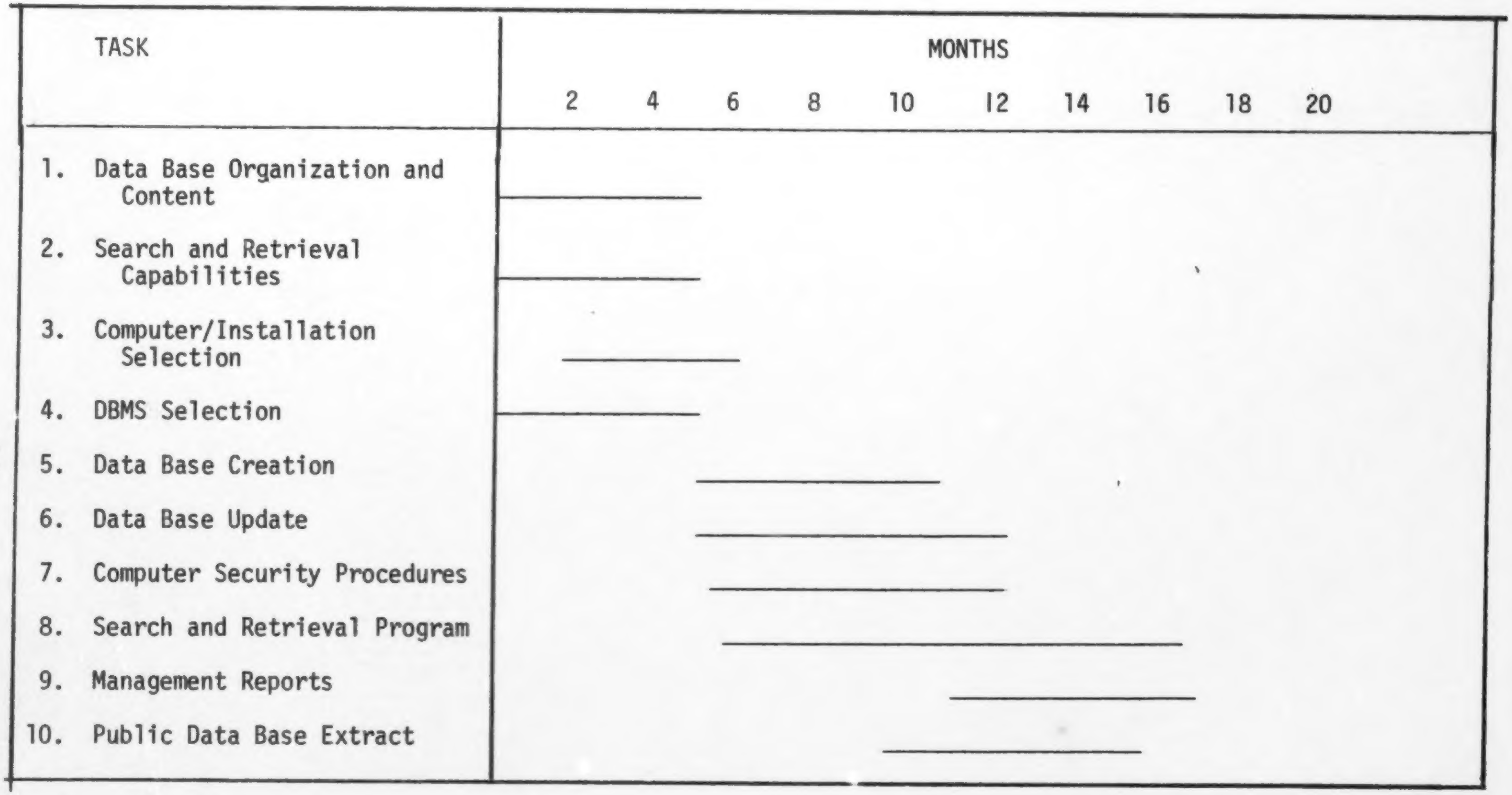
In this task, the computer and installation to process the TSCA CDS, would be selected. A decision would have to be made whether to use an installation provided by the OtS data systems contractor or an installation currently processing EPA applications, such as COMNET at Research Triangle Park. If only a single DBMS is found (Task 4) that will satisfy the TSCA CDS requirements and the DBMS is not available on any installation being considered, primary consideration should be given to making the DBMS available at one of these installations. Secondary consideration should be given to a service bureau installation that has the DBMS available. Security methods and procedures incorporated at a particular installation should play an important part in the selection process.

Task 4 - DBMS Selection

A commercial data base management system provides the user with a fully developed set of software to organize, maintain, and query a data base. The additional cost of leasing or buying a DBMS is more than offset by the reduction of system

TABLE 5-4

TSCA CDS PLAN II IMPLEMENTATION TASKS



development time and costs that can be realized by using such a package. It is therefore recommended that a survey be made of DBMS's that meet the TSCA CDS requirements defined in Task 1 and 2. If the survey indicates that the requirements could not be met by any DBMS, the data base organization, and search and retrieval requirements may be modified in order to accommodate a particular DBMS. However, the needs of OTS should not be subordinated for the sake of implementing a particular data base management system. When selecting a DBMS, primary importance should be given to the privacy and security capabilities provided by the system. If a DBMS cannot be found, the TSCA CDS organization and maintenance will have to be implemented using custom written software.

Task 5 - Data Base Creation

This task is identical to Task 6 described in Section 5.2.4.1.

Task 6 - Data Base Update

This task is identical to Task 7 described in Section 5.2.4.1.

Task 7 - Security Procedures

Security measures must be defined and developed to monitor and control unauthorized attempts to access and/or change confidential information in the TSCA CDS. Section 5.2.4.3, discusses in more detail the security considerations involved in the two alternatives.

Task 8 - Search and Retrieval Program

This task is identical to Task 9 discussed in Section 5.2.4.1.

Task 9 - Management Reports

This task is identical to Task 10 discussed in Section 5.2.4.1.

Task 10 - Public Data Base Extract

This task is identical to Task 11 discussed in Section 5.2.4.1.

Under Plan II, an initial version (Phase I TSCA reporting) of the proprietary TSCA CDS could be developed within 12 months. A fully operational version including provisions for Phase II reporting data, full search and retrieval capabilities, management reporting, and a public version of the data base could be available within 16 months. It is estimated that the 10 implementation tasks would require 110 man months of effort and cost around $450,000. This does not include the cost for computer operations. The level of effort would be divided between project managers, systems analysts, and programmers.

The computer operations costs for developing the TSCA CDS are estimated at $90,000. This includes CPU processing, storage, and terminal charges.

5.2.4.3 Security Considerations

The Mitre report and recent concerns voiced in public hearings indicated that the protection of confidential information submitted by industry from unauthorized disclosure is one of the most important consideration for OTS in implementing the TSCA Chemical Data System. Any breech of security could have a deleterious affect on a companies' competitive position in the marketplace and present EPA with serious legal problems. If EPA could show industry that the best possible security mechanisms are being employed to protect confidential information, industry would be more willing to submit confidential data under TSCA. Section 4 of this report discussed security risks posed by computer systems and mechanisms that could be used to combat these risks. The following discussion identifies the security requirements posed by the two implementation alternatives for the TSCA CDS, and recommends security mechanisms identified in Section 4 that could be implemented.

One security mechanism recommended in the Mitre report involves creating two systems; a proprietary TSCA CDS and a public TSCA CDS. The proprietary system would contain both confidential and non-confidential information and be accessed only by OTS personnel. The public system would be a subset of the proprietary system and contain only non-confidential information. The public data base would be available on a national timesharing network and could be accessed by anyone. This approach eliminates the need for one level of access control authorization. A user accessing a particular system would be either a confidential or public user. Therefore, elaborate privacy controls would not have to be implemented to protect various data elements in the data base that were declared confidential versus those that were not. Under this scheme, all data would be considered either confidential or public, and a search and retrieval program would not have to distinguish between a confidential and public user, when accessing data. This two-data-base concept is recommended for both implementation alternatives.

Security consideration played an important part in developing the two TSCA CDS implementation alternatives. Under Plan I, the proprietary TSCA CDS would be implemented using a dedicated minicomputer, while under Plan II it would be implemented using a larger scale timesharing computer. The minicomputer approach would eliminate many of the security threats discussed in Section 4. The proprietary TSCA CDS would be the only application operating on the minicomputer. The system would not be shared between confidential and non-confidential users.

Penetration of the operating system by a non-confidential users to dump a confidential data set would not have to be controlled since all the data sets and programs accessed by the operating system would be confidential. For the same reason, protection of the data sets and programs through the use of passwords would be unnecessary.

The major security requirement for the dedicated minicomputer approach is controlling and restricting user access to the system. Users must therefore be thoroughly verified before access to the system is permitted, because once access is gained, user restriction of system facilities is reduced. Authorization of relationships between system users and system resources would be unnecessary since all the system resources would be considered confidential. One means of controlling access to the system is by controlling the terminals used to access the proprietary TSCA CDS. Terminals should be located in isolated secure areas that could be locked when not in use. Terminals themselves should be provided with on-off locks as an extra measure of security. These locks should be different for each terminal and room and changed on a periodic basis.

The TSCA CDS information needs of the EPA regional offices have not yet been defined. If it is determined that the EPA regional offices require remote on-line access to the proprietary TSCA CDS information, the threat of wiretapping would pose a significant problem. As an alternative, the regional offices could be required to access the public TSCA CDS for as much information as possible and to phone in requests for confidential information to a central location within OTS. The proprietary TSCA CDS would be queried for the appropriate information which would then be sent to the regional office. If this procedure proves inadequate and the only solution is remote on-line access by EPA regional offices, a data base and/or communications encryption scheme, discussed in Section 4.3.5, should be implemented.

Within the central OTS Offices, the threat of wiretapping could be minimized if the minicomputer was installed near (i.e. 500 to 1000 feet) the OTS terminals so that a direct line (hard wired) rather than dial-up lines connected the computer and terminals. This procedure would also insure that access to the computer could only be accomplished through one of the terminals "hard-wired" to the computer. Dial-up connections allow any user with a terminal and the right telephone number to access the computer system.

The Office of Toxic Substance is considering restricting access by various users within OTS to certain data elements within the proprietary TSCA CDS to provide an even greater measure of privacy. If this option is adopted, security could be controlled by a system of subjects and objects discussed in Section 4.3.2. The subjects would be the various OTS users and the objects would be the data elements in the TSCA CDS that could be accessed by a particular OTS user. This authorization definition data would be incorporated into the on-line search and retrieval program. The OTS user would be identified to the program by a unique code or password. The user code would be matched against a set of allowable data elements for that code every time the user developed a search query. If the user specified a data element not on the allowable list, the query would not be processed. If a DBMS is utilized this capability may be part of the system package provided by the vendor. If not, the procedure will have to be coded into the search and retrieval program. This mechanism would apply to both Plan I and Plan II.

Implementing the proprietary TSCA CDS on a large scale timeshared computer poses significant security problems. As was pointed out in Section 4 there is no timesharing computer that provides complete security from unauthorized penetration. A timesharing system processes many applications and serves both the confidential user and the general non-confidential user. Since its structure is of a generalized nature, it is more prone to active threats by systems personnel and outside users than a dedicated computer system. Under Plan II, the proprietary data base system should be implemented with a verification mechanism to distinguish between the general user and the privileged confidential OTS user. The data files and computer programs applicable to the TSCA CDS should be protected by a system of passwords. The executable and source versions of the computer programs should be stored in private libraries that are password protected instead of a common library shared by many general users. The computer system selected should provide the capability of establishing authorization relationships between users, terminals, programs and peripherals as discussed in Section 4.3.2. This will isolate proprietary TSCA CDS processing and resources from other users in the timesharing system. Many commercial data base management systems provide this type of security and should receive prime consideration when selecting a DBMS.

The computer system selected should possess a virtual storage operating system in order to provide a protection mechanism for data areas and areas for executable code. Since the Plan II computer installation would probably not be near the central OTS offices, on-line terminal access to the computer could not be "hard-wired" as in Plan I. A dial-up method would therefore have to be employed. Since this is required for the central OTS offices, regional EPA offices would also be permitted on-line access. Since the TSCA CDS would be in a timesharing environment and data would be transmitted over general communication lines a data base encryption scheme should be employed. A discussion of encryption is given in Section 4.3.5. There are some data base management system packages, such as Adabas, that provide the necessary encryption requirements. The National Bureau of Standards (NBS) provides a federal encryption algorithm. Numerous other algorithms are provided by commercial vendors. Terminals are available that can both send and receive encrypted information. These mechanisms, while not fool-proof, do provide an added measure of security and should receive prime consideration when making implementation decisions.

The most serious threats to the proprietary TSCA CDS would come from users attempting to browse through confidential information rather than from users attempting to alter certain information. Therefore, on-line updating does not pose as many security requirements as on-line search and retrieval does. However, for security purposes a batch update procedure, should be used instead of an on-line procedure, to reduce the chance of unauthorized penetration.

Software for both implementation plans should be developed according to the procedures discussed in Section 4.3.6. The slaient points are that software should be developed in-house as much as possible, written in a higher-level language, modularly designed, audited line-by-line, and updated in a strictly controlled manner.

5.2.4.4 Advantages and Disadvantages of Each Plan

The advantages and disadvantages of Plan I and II are compared in order to identify implementation trade-offs and considerations. Table 5-5 presents a summary of these strengths and weaknesses.

Under Plan I, the TSCA CDS would be processed using a dedicated minicomputer. The major advantage to this plan is the amount of security and privacy such a system would provide. Although not eliminated, security requirements would be greatly minimized compared to a large scale timesharing system. The major security requirement would be the verification of a users identity and access authorization. Since the system would be dedicated to the needs of OTS, it could be configured to provide the optional processing environment according to OTS requirements. OTS would be able to control the system by specifying the type and number of computer programs that are to execute at any given time in

TABLE 5-5

A COMPARISON OF TSCA CDS IMPLEMENTATION PLANS

PLAN	ADVANTAGES	DISADVANTAGES
I Dedicated Minicomputer	OTS control of system. Security requirements minimized. Computer system dedicated to a single application. EPA approved minicomputer procurement policy.	Limited number of DBMS packages available. Limited amount of vendor software. Programming languages are not full versions. Less utilization of computer system. Limited number of experienced personnel. Installation required.
II Large scale Timesharing computer	System immediately available. Abundant vendor supplied software. Wide selection of DBMS packages. Computer system utilized more fully.	Protection of confidential data. Contention with other users. System not controlled by OTS.

the system. EPA has already established a minicomputer policy which should make procurement much easier. The standard EPA minicomputer configuration would very likely meet the OTS processing requirements. The minicomputer could be installed close to the central OTS offices so that terminals could be "hard-wired" to the computer to provide maximum terminal-user security. The major disadvantage with Plan I concerns computer software. Only a limited number of data base management packages are available for minicomputers and these packages are quite often subsets or watered-down versions of the DBMS packages available on large-scale computers. Computer language compilers, such as COBOL or Fortran, are often not the full versions found on the larger machines. Vendor supplied software utility programs are not as abundant on minicomputers as they are on larger machines. These utilities include text editors, file dumps, print routines, copy routines, etc. which makes the sytem development process for programmers much easier. For minicomputer systems, much of the utility software must therefore be coded by the programming staff. Since the minicomputer field is relatively new and not as widely used as larger computers, there is a limited number of qualified personnel for system development and operation. However, this is changing rapidly as the field expands. Because of these factors, systems development work usually takes more time to complete and often costs more. Since the minicomputer would be dedicated to a single application, full computer utilization may not be realized. The processing requirements of OTS may necessitate only a single or double shift, leaving the computer idle for one or two shifts. Processing during any shift is limited to the needs of OTS. Under Plan I a computer would have to be selected and purchased, a site would have to be selected and prepared, and the computer would have to be installed before operation could begin. As indicated by Table 5-3 this could take as long as 12 months.

Plan II involved implementing the TSCA CDS on a large-scale timesharing system. One advantage to this approach is that a wide selection of data base management packages are available on these machines. There is also abundant vendor supplied software such as compilers and utility programs that can increase programmer productivity. Since EPA currently utilizes large-scale timesharing systems, access to computer processing facilities would be available immediately. Since many users share the computer, OTS would be billed only for the time and resources it actually uses.

The major disadvantage to Plan II involves the protection of confidential information. Since the computer is shared by many users, the opportunity and threat of unauthorized access is much greater. As was pointed out in Section 4 there is virtually no timesharing system that can adequately protect confidential data. Some additional security requirements under Plan II, that are not needed under Plan I include:

- Operating system protection
- General user versus confidential user identification authorization checking
- Protection of data files and data bases
- Protection of computer program libraries
- Access control to system resources

Since the computer facility will probably be some distance from the OTS offices, on-line terminal access to the TSCA CDS would be through dial-up procedures. This increases the chance of wiretapping and increases the requirements for verifying a users identity. Under a timesharing environment, OTS must contend with other system users for computer processing and system resources.

5.3 OTS Interaction With CSIN

The proposed Chemical Substances Information Network is a comprehensive integrated data network that is responsive to the needs of OTS in meeting its obligations under the Toxic Substances Control Act. The network provides OTS the means to collect and store information related to TSCA share this information with other agencies and the general public, conduct hazard identification and hazard assessment of chemical substances affecting human health and the environment, and regulate these chemical substances that pose unreasonable risk to human health or the environment.

Since EPA/OTS will be a primary user of the CSIN, this Section discusses how the Office of Toxic Substances and its related TSCA activities would interact with the Chemical Substances Information Network. Various functional areas within OTS are identified and their activities discussed. Associated with the activities for each functional area is a discussion of the identity and uses of CSIN component data bases that might be required. A conceptual operational-level description of the OTS/CSIN interface is presented for two periods of time. The short-term description reflects the first 5 years, while the long-term description reflects years 5 through 10.

5.3.1 OTS Functions Related to CSIN Data Bases

Figure 5-7 illustrates a conceptual view of a hierarchical relationship between various functional areas within OTS. The three major areas within OTS include:

- Information and Program Integration
- Testing and Evaluation
- Chemical Control

The Information and Program Integration area is responsible for collecting, maintaining, and disseminating information related to the Toxic Substances Control Act. This area is further divided into three functional groups which include:

- Information
- Problem Identification
- Coordination and Liaison

The testing and evaluation area is responsible for assessing possible risks of chemical substances to human health and the environment. This area is composed of three functional groups:

- Hazard assessment
- Health effects
- Ecological effects

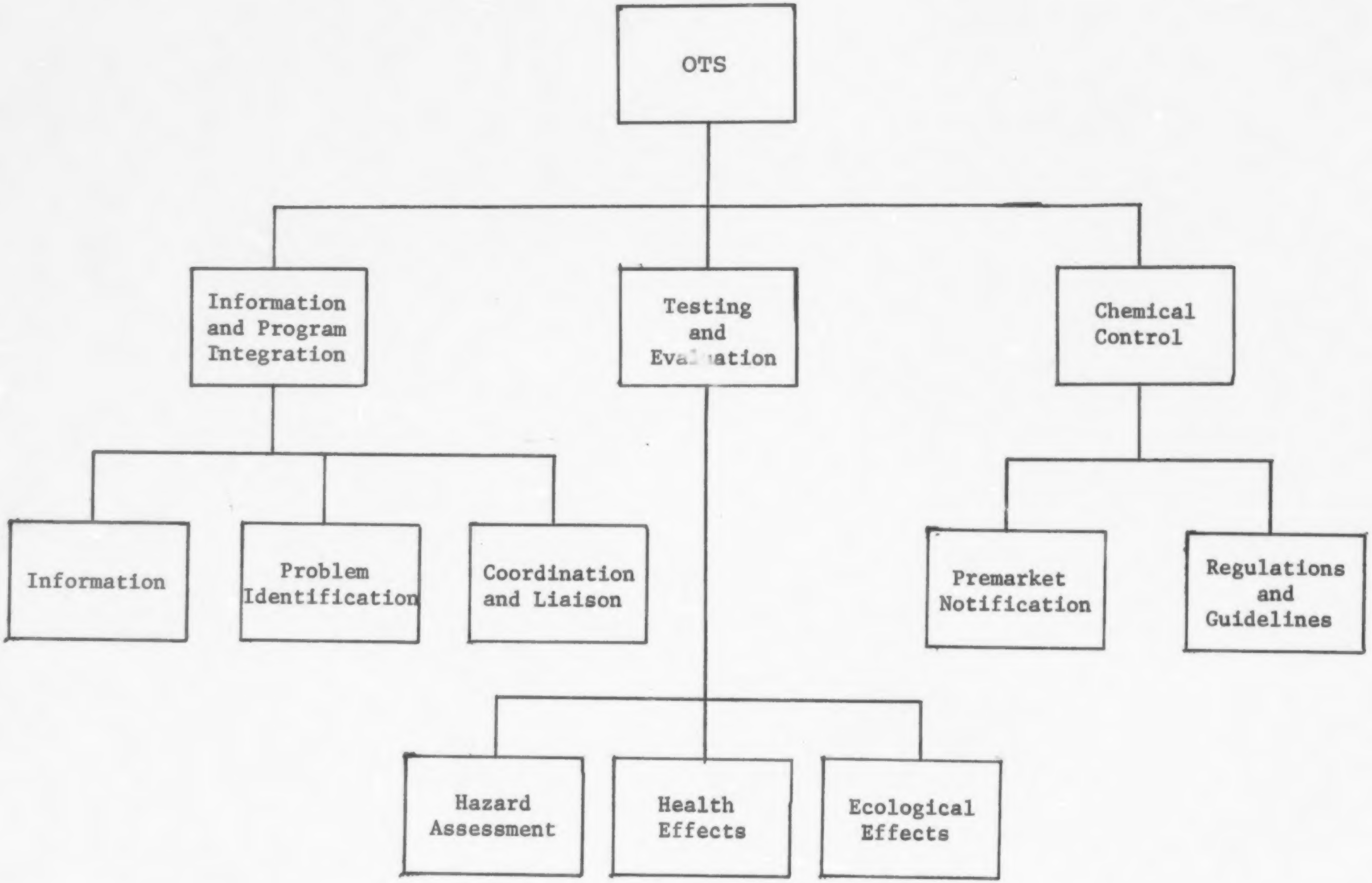
The Chemical Control area is responsible for using TSCA regulatory authorities to control new chemical substances and existing chemical substances. The two functional groups that make up the Chemical Control area include:

- Premarket Notification
- Regulations and guidelines

A more detailed discussion is presented concerning the activities of these eight function groups within OTS and how they interact with component files of the CSIN identified in the Mitre report. The activities of these functional groups are discussed according to OTS involvement with four major situations. These four situations include:

FIGURE 5-7

FUNCTIONAL AREAS WITHIN OTS


OTS
Information and Program Integration
Testing and Evaluation
Chemical Control
Information
Problem Identification
Coordination and Liaison
Premarket Notification
Regulations and Guidelines
Hazard Assessment
Health Effects
Ecological Effects

- General reporting
- Premarket notification
- Notification of possible risk
- Identification/priorization of chemical substances

General reporting includes developing TSCA reporting guidelines and requirements for information concerning site specific production, premarket notification, exposure, users, byproducts, impurities, adverse reactions, health records, and chemical testing. It also includes collecting, storing, and disseminating this information. Premarket notification is concerned with reviewing and controlling new chemical substances that are introduced into the Commerce. It includes developing premarket notification requirements, reviewing each premarket notification, assessing the risk to human health and the environment, and taking regulatory steps to control the new chemical substance. Notifications of possible risks are received from citizens, organizations, other agencies, etc. and must be processed by OTS. This includes reviewing and assessing the possible risk, conducting any required tests, making decisions concerning the control of chemical substances involved, and making the decisions known to interested parties. Identification/priorization is concerned with selecting chemical substances or classes of chemical substances from the chemicals in Commerce (TSCA Inventory) that pose a substantial risk to human health and the environment. The selected chemicals detailed information may be gathered and specific tests conducted through TSCA reporting authorities.

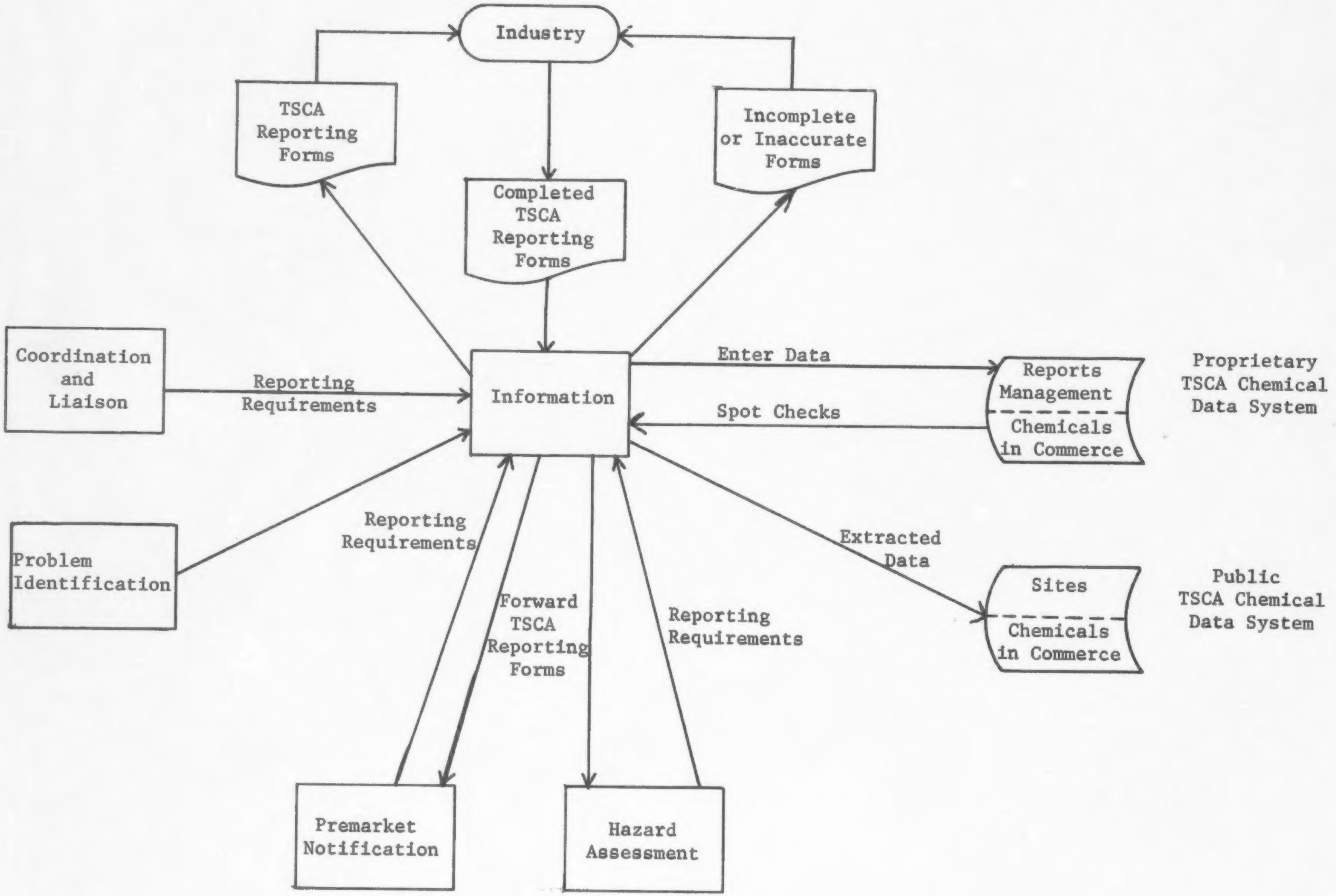
5.3.1.1 Information

EPA has the authority under TSCA to require industry to provide certain information concerning chemical substances. The information group within OTS would be responsible for collecting and storing this information. Figure 5-8 illustrates some of the activities within the Information group.

This group would be responsible for designing the actual reporting forms. The forms that already exist or are being planned include site specific production, premarket notification, uses, occupational exposure, byproducts, impurities, adverse reactions and health studies. Depending on the reporting form, this group would receive input for design specifications from other groups such as Premarket Notification, Hazard Assessment, and Coordination and Liaison. The information group would be responsible for distributing these forms to the appropriate parties and act as a focal point for receiving all forms submitted to EPA.

FIGURE 5-8

OTS INFORMATION GROUP


Industry
TSCA Reporting Forms
Completed TSCA Reporting Forms
Incomplete or Inaccurate Forms
Coordination and Liaison
Reporting Requirements
Information
Enter Data
Spot Checks
Reports Management
Chemicals in Commerce
Proprietary TSCA Chemical Data System
Problem Identification
Reporting Requirements
Extracted Data
Sites
Chemicals in Commerce
Public TSCA Chemical Data System
Forward TSCA Reporting Forms
Reporting Requirements
Premarket Notification
Hazard Assessment

To store all the information submitted, the Information group would be involved with the design and development of the proprietary TSCA Chemical Data System. The data base would contain information about the reports submitted to OTS (Reports Management file) and the data contained in those reports (Chemicals in Commerce file). The proprietary data base would contain both confidential and non-confidential information.

All request for TSCA reporting information by other groups would be processed through the Information group. As TSCA reports are received, they would be logged into the Reports Management file. The forms would be edited for accuracy and completeness. Forms that did not pass the edit would be returned to the submitter for correction. Forms that passed the edit would have their data entered into the Chemicals in Commerce file. In the case of test results information, acutal data would not be entered into the system. Instead, an indication that the test has been conducted and where the data can be found would be entered into the system. Copies of certain forms submitted would be forwarded to groups such as Premarket Notification and Hazard Assessment for review and analysis.

The Reports Management System would indicate what reports should be submitted, by what parties and in what time frame. The Information group would perform periodic spot checks on TSCA respondents. The Reports Management file would be searched to determined what parties have not submitted required reports, and what parties are past their submission due date for certain reports.

The Information group would also be responsible for confidential information related to the TSCA CDS. This includes specifying how confidentiality claims are to be declared on the reporting forms, and protecting the privacy of confidential data on the proprietary TSCA CDS. The proprietary data base would be designed to distinguish between confidential and non-confidential information, and prevent unauthorized users from accessing the data base. Only OTS personnel would be allowed to access the proprietary TSCA CDS. Security procedures would be defined, implemented, and monitored by the Information group to insure the integrity of the proprietary TSCA CDS. Information obtained under TSCA would be made available for public use as quickly as possible, through the public TSCA CDS.

The Information group would be responsible for the design of the public data base, extracting the non-confidential information from the proprietary TSCA CDS, and updating the public data base on a periodic basis. By providing a public TSCA CDS, OTS will enable interested parties to solicit information directly from the public data base, thus shifting the burden of answering questions away from OTS personnel.

5.3.1.2 Problem Identification

The Problem Identification group would be responsible for responding to chemical problems that have been brought to the attention of EPA and that might pose serious risks to human health and/or the environment. TSCA provides several mechanisms by which notification of substantial risk can be made. These mechanisms include citizens petitions, risks uncovered by test results, and notification of substantial risk by industry. Other federal or state agencies, or interagency groups may also notify OTS of a substantial risk. The Problem Identification Group would respond to these notifications within a reasonable time frame depending on the urgency of the problem. Figure 5-9 illustrates some of the activities within the Problem Identification group.

As notifications are received by the Problem Identification group, they would be logged into the Reports Management file of the TSCA CDS through the Information group to provide a tracking capability. The Problem Identification group would include a team of technical specialists that would be available on an on-call basis to assess risks and provide specifications for prompt and effective corrective actions to urgent problems.

The chemicals or classes of chemicals specified in the notification must be identified and the proper priority of responsibility determined. For instance, if the chemical was under the jurisdiction of FDA, it would be referred to that agency for appropriate assessment and action. The Chemicals In Commerce file of the TSCA CDS could be accessed to determine the amount in Commerce, uses, and exposure in making a preliminary assessment of the chemical substance.

If more information is needed, the Chemical Structure/Nomenclature System could be accessed to determine other data bases containing additional information and/or other chemical substances that are similar to chemical substance posing the risk. In addition, the Chemical Data Bases Directory may be accessed to determine what data bases contain a certain type of information desired.

FIGURE 5-9

OTS PROBLEM IDENTIFICATION GROUP

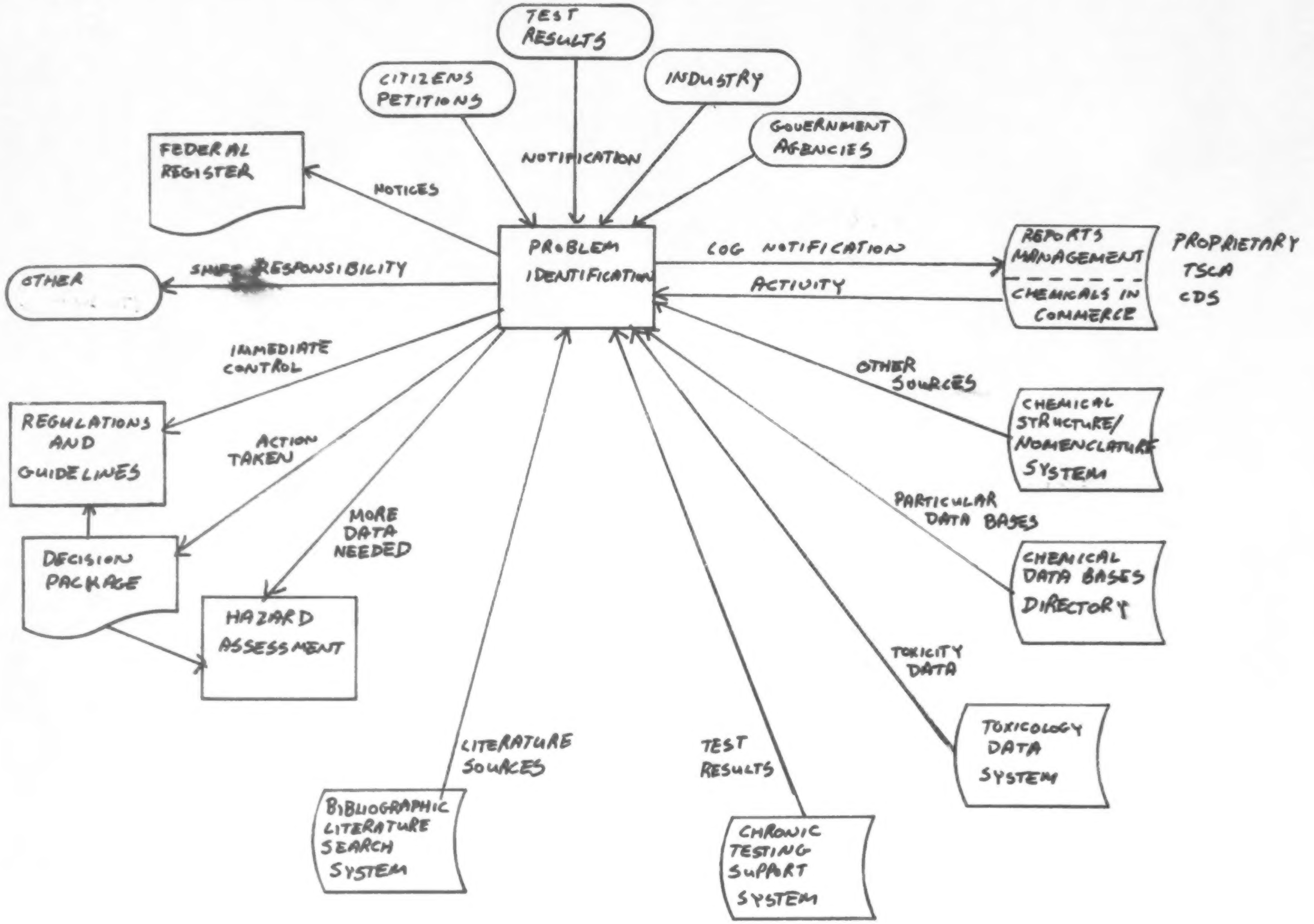
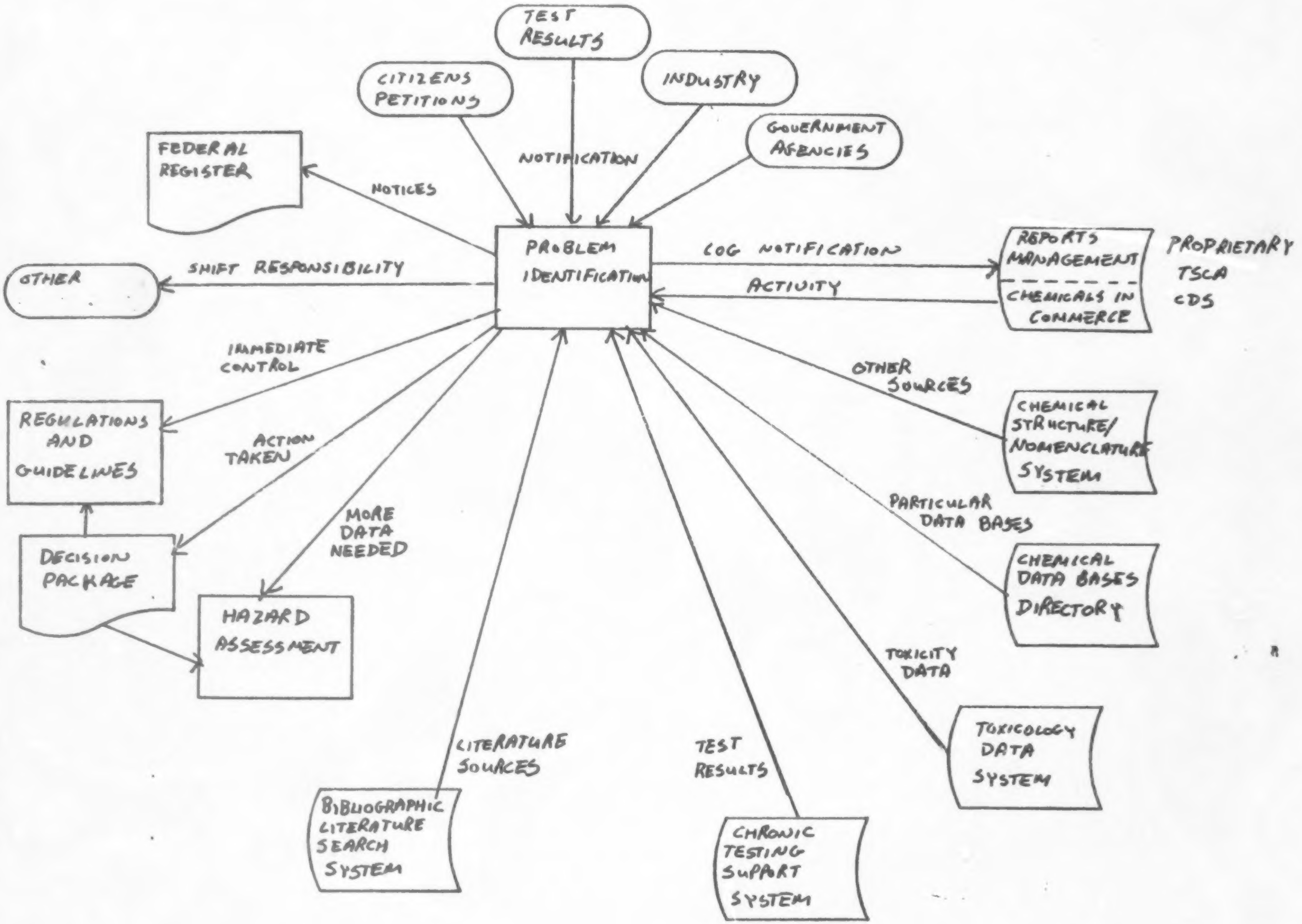

TEST RESULTS
CITIZENS PETITIONS
INDUSTRY
GOVERNMENT AGENCIES
NOTIFICATION
FEDERAL REGISTER
NOTICES
PROBLEM IDENTIFICATION
LOG NOTIFICATION
ACTIVITY
REPORTS MANAGEMENT
CHEMICALS IN COMMERCE
PROPRIETARY
TSCA
CDS
OTHER
SHIFT RESPONSIBILITY
IMMEDIATE CONTROL
REGULATIONS AND GUIDELINES
ACTION TAKEN
DECISION PACKAGE
MORE DATA NEEDED
HAZARD ASSESSMENT
OTHER SOURCES
CHEMICAL STRUCTURE/ NOMENCLATURE SYSTEM
PARTICULAR DATA BASES
CHEMICAL DATA BASES DIRECTORY
TOXICITY DATA
TOXICOLOGY DATA SYSTEM
LITERATURE SOURCES
BIBLIOGRAPHIC LITERATURE SEARCH SYSTEM
TEST RESULTS
CHRONIC TESTING SUPPORT SYSTEM

FIGURE 5-9

OTS PROBLEM IDENTIFICATION GROUP


TEST RESULTS
CITIZENS PETITIONS
INDUSTRY
GOVERNMENT AGENCIES
NOTIFICATION
FEDERAL REGISTER
NOTICES
PROBLEM IDENTIFICATION
LOG NOTIFICATION
ACTIVITY
REPORTS MANAGEMENT
CHEMICALS IN COMMERCE
PROPRIETARY
TSCA
CDS
OTHER
SHIFT RESPONSIBILITY
IMMEDIATE CONTROL
REGULATIONS AND GUIDELINES
ACTION TAKEN
DECISION PACKAGE
MORE DATA NEEDED
HAZARD ASSESSMENT
OTHER SOURCES
CHEMICAL STRUCTURE/ NOMENCLATURE SYSTEM
PARTICULAR DATA BASES
CHEMICAL DATA BASES DIRECTORY
TOXICITY DATA
TOXICOLOGY DATA SYSTEM
LITERATURE SOURCES
BIBLIOGRAPHIC LITERATURE SEARCH SYSTEM
TEST RESULTS
CHRONIC TESTING SUPPORT SYSTEM

The Toxicology Data System would provide detailed information. The Chronic Testing Support System may be accessed to determine what tests have been conducted on the chemical substance and the results of those tests. If these data bases fail to provide enough information, literature sources may have to be investigated. The Bibliographic Literature Search System would provide a convenient mechanism for determining what available literature is applicable to the chemical substance.

Based on the investigation conducted by the Problem Identification group a decision package would be developed indicating the recommended action to be taken. There are several actions that could be recommended by this group. The investigation results could indicate that there is no risk to human health or the environment, in which case no action would be taken. If the investigation indicated that insufficient data was available, additional testing or monitoring may be recommended. In this case, the investigation would be continued by the Hazard Assessment group which would specify the necessary tests and assess the results of these tests. The investigation may conclude that immediate control of the chemical substance is necessary. In this case the Regulations and Guidelines groups would specify the type of control to be imposed. This could include banning or limiting the manufacture, processing or use of the chemical substance, specifying warning labels, or specifying disposal or processing methods.

5.3.1.3 Coordination and Liaison

The Coordination and Liaison group would be responsible for handling inquiries of parties interested in data collected under TSCA and stored in the TSCA Chemical Data System. Figure 5-10 illustrates some of the activities associated with the Coordination and Liaison group.

OTS has a dual role as far as the information collected under TSCA is concerned. Since a portion of the data required to adequately assess potential risks to human health and the environment would be declared confidential by industry, OTS must protect that data from unauthorized disclosure. On the other hand, OTS has the responsibility of making as much of the TSCA information as possible available for public use. To provide the maximum dissemination of information, the TSCA Chemical Data System would consist of two data bases, i.e. a proprietary data base and a public data base. This procedure will eliminate the need of having all requests for information go through OTS. Instead, if an interested party required only

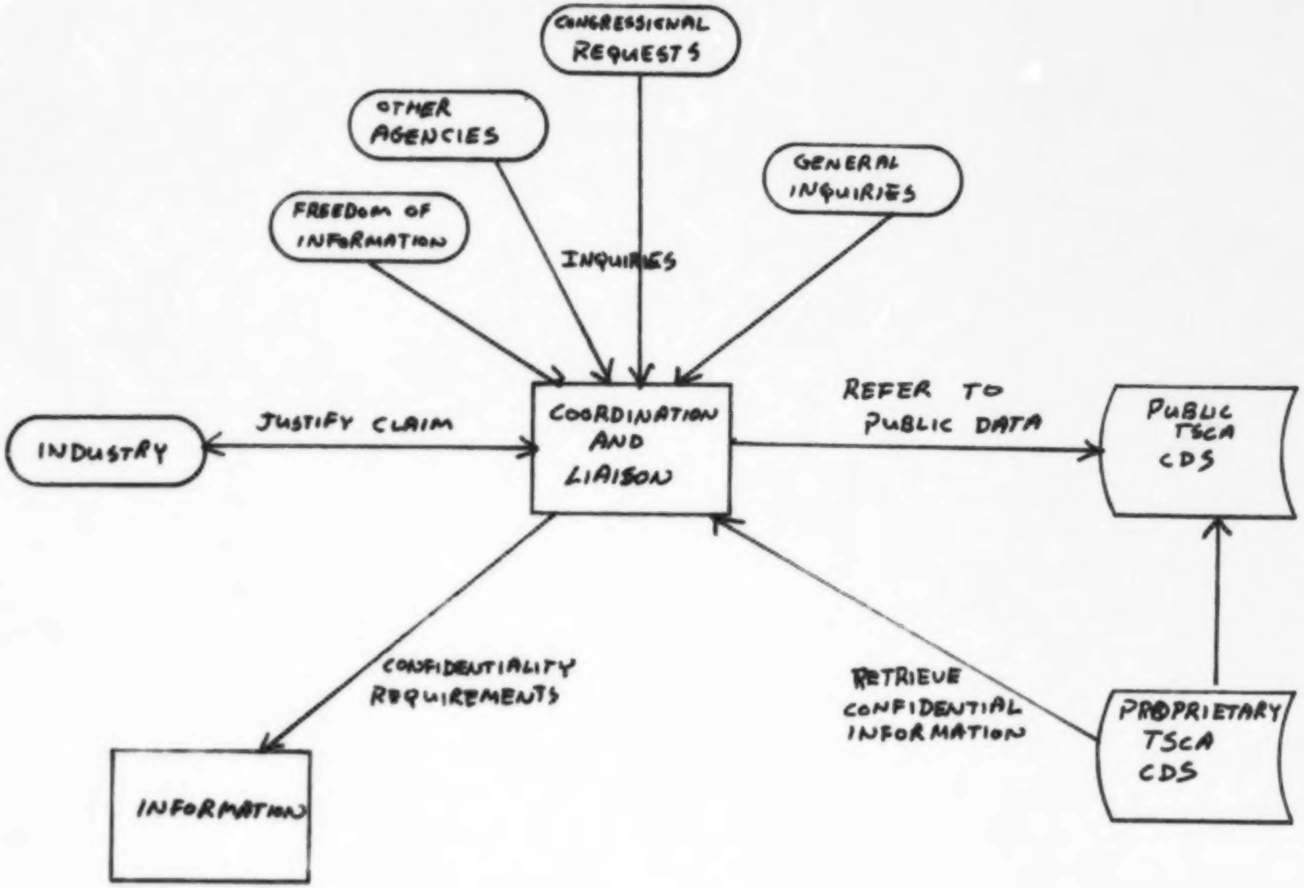

CONGRESSIONAL REQUESTS
OTHER AGENCIES
GENERAL INQUIRIES
FREEDOM OF INFORMATION
INQUIRIES
INDUSTRY
JUSTIFY CLAIM
COORDINATION AND LIAISON
REFER TO PUBLIC DATA
PUBLIC TSCA CDS
CONFIDENTIALITY REQUIREMENTS
RETRIEVE CONFIDENTIAL INFORMATION
PROPRIETARY TSCA CDS
INFORMATION

FIGURE 5-10

OTS COORDINATION AND LIAISON GROUP

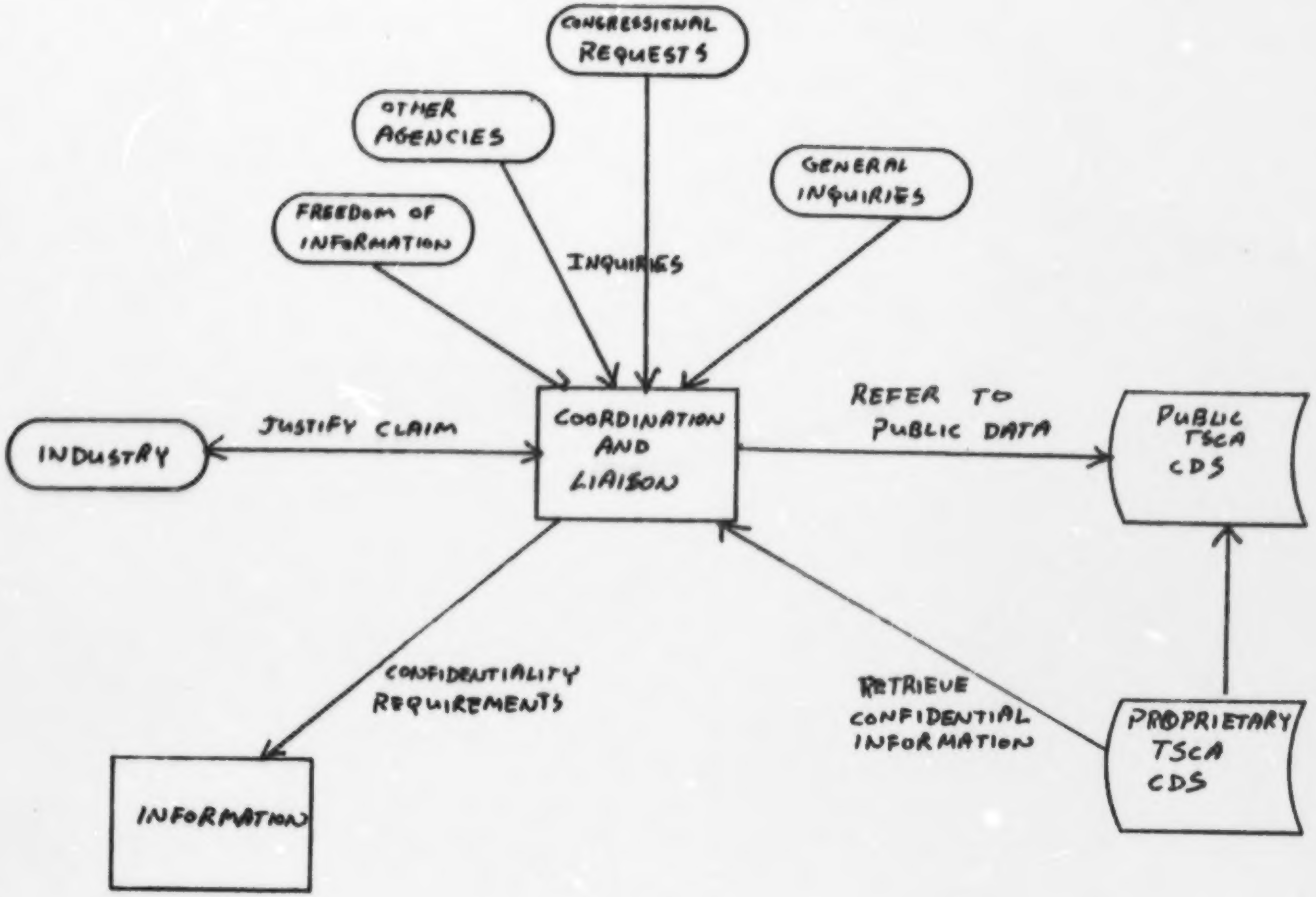

CONGRESSIONAL REQUESTS
OTHER AGENCIES
FREEDOM OF INFORMATION
GENERAL INQUIRIES
INQUIRIES
INDUSTRY
JUSTIFY CLAIM
COORDINATION AND LIAISON
REFER TO PUBLIC DATA
PUBLIC TSCA CDS
CONFIDENTIALITY REQUIREMENTS
INFORMATION
RETRIEVE CONFIDENTIAL INFORMATION
PROPRIETARY TSCA CDS

FIGURE 5-10

OTS COORDINATION AND LIAISON GROUP

non-confidential information, they would be directed to the public TSCA CDS. The Coordination and Liaison group would handle only requests for confidential information not available on the Public TSCA CDS.

Since the Information group would be responsible for designing and making the public data base available, the Coordination and Liaison group would provide input to this group concerning what data elements could be considered confidential and how confidential information could be consolidated or integrated in order to prevent actual identification. The Coordination and Liaison group would also be responsible for making the public aware of the existence of the Public TSCA CDS and how it could be used to obtain TSCA data.

Inquiries would mostly arise as a result of Freedom of Information Act requests (F.O.I.), congressional requests, or requests from other agencies such as CPSC, FDA, or OSHA, that agree to share their respective data. Each request would be analyzed to determine its validity. Invalid requests would be rejected and returned to the submitter with an appropriate explanation. Before data, designated or confidential is released in response to valid requests, the originator of the data would be given an opportunity to justify the claim of confidentiality, and seek judicial relief if the decision does not agree with OTS. The appropriate confidential data would be extracted from the proprietary TSCA CDS and may be masked in some cases to destroy the actual link to originator.

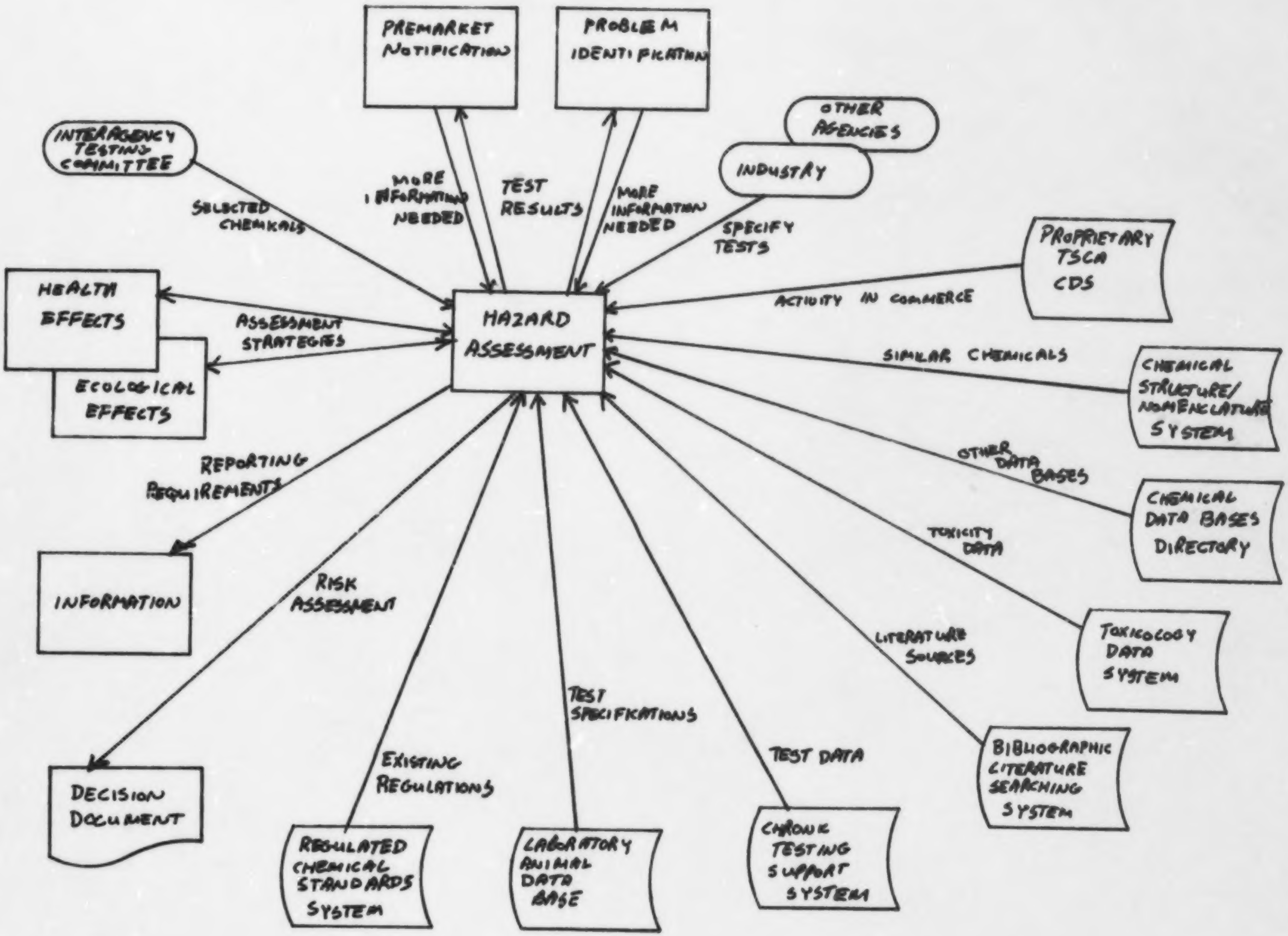
5.3.1.4 Hazard Assessment

The Hazard Assessment group has broad responsibility over specifying tests and analyzing test results in order to assess the potential risk to human health and the environment posed by certain chemical substances. Figure 5-11 illustrates some of the activities associated with the Hazard Assessment group. In order to conduct risk or hazard assessments associated with toxic substances, timely and reliable test data is needed on a wide range of chemicals. Such test data would be used to support regulatory actions and identify problem areas that require more attention.

Currently, there is no agreed upon technique or procedure for conducting environmental risk assessments. TSCA provides OTS with the ability to develop general testing requirements and protocols for categories of chemical substances that would gain universal acceptance.

FIGURE 5-11

OTS HAZARD ASSESSMENT GROUP


PREMARKET NOTIFICATION
PROBLEM IDENTIFICATION
INTERAGENCY TESTING COMMITTEE
OTHER AGENCIES
INDUSTRY
MORE INFORMATION NEEDED
TEST RESULTS
MORE INFORMATION NEEDED
SPECIFY TESTS
SELECTED CHEMICALS
PROPRIETARY TSCA CDS
ACTIVITY IN COMMERCE
HEALTH EFFECTS
ECOLOGICAL EFFECTS
ASSESSMENT STRATEGIES
HAZARD ASSESSMENT
SIMILAR CHEMICALS
CHEMICAL STRUCTURE/ NOMENCLATURE SYSTEM
OTHER DATA BASES
CHEMICAL DATA BASES DIRECTORY
REPORTING REQUIREMENTS
TOXICITY DATA
INFORMATION
RISK ASSESSMENT
TOXICOLOGY DATA SYSTEM
LITERATURE SOURCES
TEST SPECIFICATIONS
TEST DATA
BIBLIOGRAPHIC LITERATURE SEARCHING SYSTEM
EXISTING REGULATIONS
DECISION DOCUMENT
REGULATED CHEMICAL STANDARDS SYSTEM
LABORATORY ANIMAL DATA BASE
CHRONIC TESTING SUPPORT SYSTEM

The Hazard Assessment group along with the Premarket Notification, Problem Identification, and Information groups, and the Interagency Testing Committee (ITC) would provide input in determining testing requirements related to TSCA which include:

- The format for submitting test data
- The requirements for determining specific effects of specific chemicals
- The test requirements for selected categories of new and existing chemicals

The Hazard Assessment group would process requests for more detailed information relating to risk assessment from several sources. During the premarket review procedure for new chemicals entering Commerce, the Premarket Notification group may conclude after a review that more intensive testing is required before full manufacture can commence. In this case the Hazard Assessment group would be made aware of what type of data is needed and the type of risk assessment is to be made. The Problem Identification group, in a similar way, may conclude after an initial review of existing data that additional information must be obtained and evaluated before a crediable response to a risk notification can be made. The Interagency Testing Committee will select up to 50 priority chemicals from the universe of chemicals in Commerce that would require additional testing so that meaningful evaluations can be made concerning possible risk. The Hazard Assessment group would be responsible for specifying the required tests and conducting a risk assessment for these selected chemicals.

In order to conduct a risk assessment, the Hazard Assessment group would be concerned with the following activities:

- Determining what data is needed for a risk assessment
- Surveying current sources of information for relevant data
- Specifying tests to be conducted for data that does not exist
- Monitoring the testing procedure

Data that may be needed to conduct risks assessments include:

- Physical and chemical properties
- Source and exposure information

- Environmental behavior and fate
- Human effects
- Effects on plants, animals and soil.

Once a determination is made concerning what data is needed, appropriate data bases could be searched to locate the information. The Chemical Data Bases Directory would be accessed to identify sources of specific types of data. The TSCA CDS could be accessed to obtain source and exposure data such as production, uses, byproducts, impurities, and occupational exposure. Physical and chemical properties could be found in the Chemical Structure/Nomenclature System and the Toxicology Data System. The TDS could also be used to obtain environmental behavior and fate information. The Chronic testing support system could be used to obtain information on Human Effects and effects on plants, animals, and soil. Data relevant to court opinions relating to a particular substance or risk could be obtained from the Regulated Chemical Standards System.

Associated with the Hazard Assessment group are the Health Effects and Ecological Effects groups that will provide technical support in developing assessment strategies, obtaining detailed information, and conducting special tests. If required data is not currently available, the Hazard Assessment group would specify the test data required and the parties responsible for conducting the tests. The Laboratory Animal Data Base could be used to determine tests to be conducted using animals. Private industry would be responsible for conducting many tests and preparing risk assessments under TSCA. Other agencies or the academic field will also contribute to the testing requirements. In some instances the data required could be obtained through the general reporting provisions of TSCA. In this case the Information group would be responsible for establishing the reporting requirements, collecting the data, and entering it into the TSCA CDS.

The Hazard Assessment group would monitor the testing procedure until it is completed. Once the required data is obtained, a risk assessment in the form of a decision document will be prepared and sent to the appropriate party (Interagency Testing Committee, Premarket Notification, or Problem Identification).

5.3.1.5 Health Effects

The Health Effects group will consist of technical experts in the area human and animal health that would aid the Hazard Assessment group in defining test requirements and developing risk assessments. The activities associated with this group are illustrated in Figure 5-12. The Hazard Assessment group will have overall responsibility for risk assessments. The Health Effects group would be utilized by the Hazard Assessment group to conduct in-depth reviews and evaluation of problems concerning health effects. This group would also provide input to the Hazard Assessment group concerning the testing requirements needed to obtain more information.

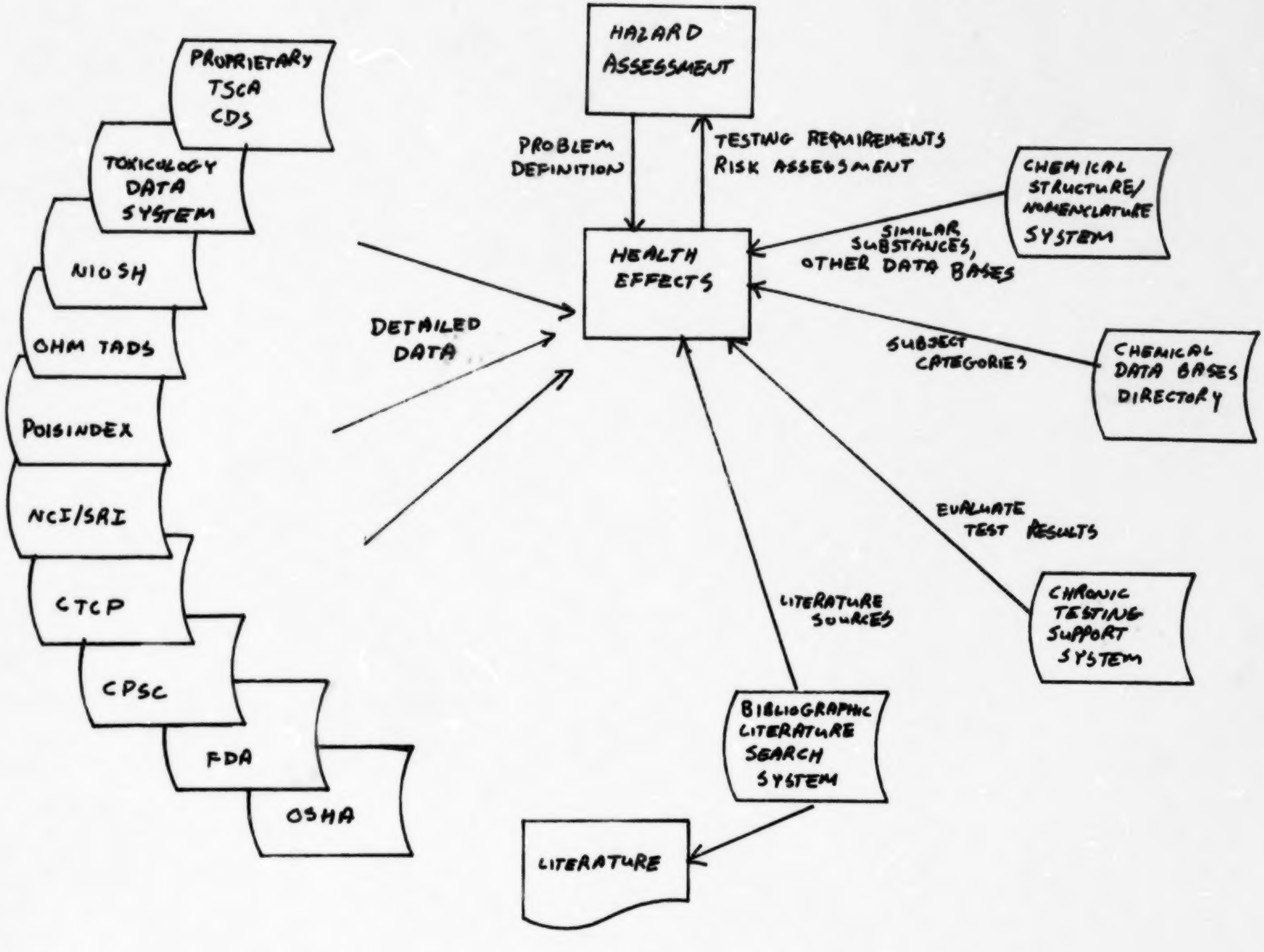
The Health Effects group would act mainly in a research capacity and be involved in long-term in-depth review and analysis of problem associated with toxic substances. The group would be deeply involved with the priority chemicals selected by the Interagency Testing Committee. Since the group would be concerned with very detailed information, access to a wide variety of data bases would be required.

The Chemical Structure/Nomenclature System could be used to identify similar compounds to the one under review in order to determine possible corresponding risks. The data base could also be used to identify other data bases that contain additional information about the substance. The Chemical Data Bases Directory could be used to determine what data base contain information about a particular subject category under review, or to determine how an unknown data base can be accessed. The Chronic Testing Support System could be accessed to evaluate test results collected for a particular chemical. The Bibliographic Literature Search System could be accessed to determine what journals contain information relevant to the chemical under review.

Data bases, applicable to health effects, that could be accessed to obtain relevant data include the TSCA CDS, Toxicology Data System, NIOSH, Oil and Hazardous Materials, POISINDEX, Research Program of Chemicals that Impact Man (SRI), and Clinical Toxicology of Commercial Products. The files of the Consumer Product Safety Commission FDA, and OSHA might also be accessed

FIGURE 5-12

OTS HEALTH EFFECTS GROUP


HAZARD ASSESSMENT
PROBLEM DEFINITION
TESTING REQUIREMENTS
RISK ASSESSMENT
HEALTH EFFECTS
CHEMICAL STRUCTURE/ NOMENCLATURE SYSTEM
SIMILAR SUBSTANCES, OTHER DATA BASES
SUBJECT CATEGORIES
CHEMICAL DATA BASES DIRECTORY
EVALUATE TEST RESULTS
CHRONIC TESTING SUPPORT SYSTEM
LITERATURE SOURCES
BIBLIOGRAPHIC LITERATURE SEARCH SYSTEM
LITERATURE
DETAILED DATA
PROPRIETARY TSCA CDS
TOXICOLOGY DATA SYSTEM
NIOSH
OHM TADS
POISINDEX
NCI/SRI
CTCP
CPSC
FDA
OSHA

5.3.1.6 Ecological Effects

The Ecological Effects group would consist of technical experts in the area of environmental effects that would aid the Hazard Assessment group in defining test requirements and developing risk assessments. Figure 5-13 illustrates some of the activities within this group. The Ecological Effects group shares the same type of association with the Hazard Assessment group as the Health Effects group. They would carry on the same types of activities except that this group is concerned with the environmental effects and not health effects. The group would act mainly in a research capacity and be involved in long-term in-depth review and analysis of problems associated with toxic substances. The group would also be deeply involved with the priority chemicals selected by the Interagency Testing Committee. Since the group would be involved in in-depth ad hoc reviews they will need access to a wide variety of data bases.

The Chemical Structure/Nomenclature System could be used to identify similar compounds to the one under review in order to determine similar risks. The data base could also be used to identify other data bases that contain additional information about the substance. The Chemical Data Bases Directory could be used to determine what data bases contain information about a particular subject category under review, or to determine how an unknown data base can be accessed. The Chronic Testing Support System could be accessed to evaluate test results collected for a particular chemical. The Bibliographic Literature Search System could be accessed to determine what journals contain information relevant to the chemical under review.

Data bases applicable to ecological effects that might be accessed for detailed data include the TSCA CDS, Toxicology Data System, Research Program of Chemicals that Impact Man, Solid Waste Information Retrieval System.

5.3.1.7 Premarket Notification

The Premarket Notification group would be responsible for determining the environmental acceptability of all new chemicals that are introduced into Commerce. Figure 5-14 illustrates some of the activities associated with this group.

FIGURE 5-13

OTS ECOLOGICAL EFFECTS GROUP

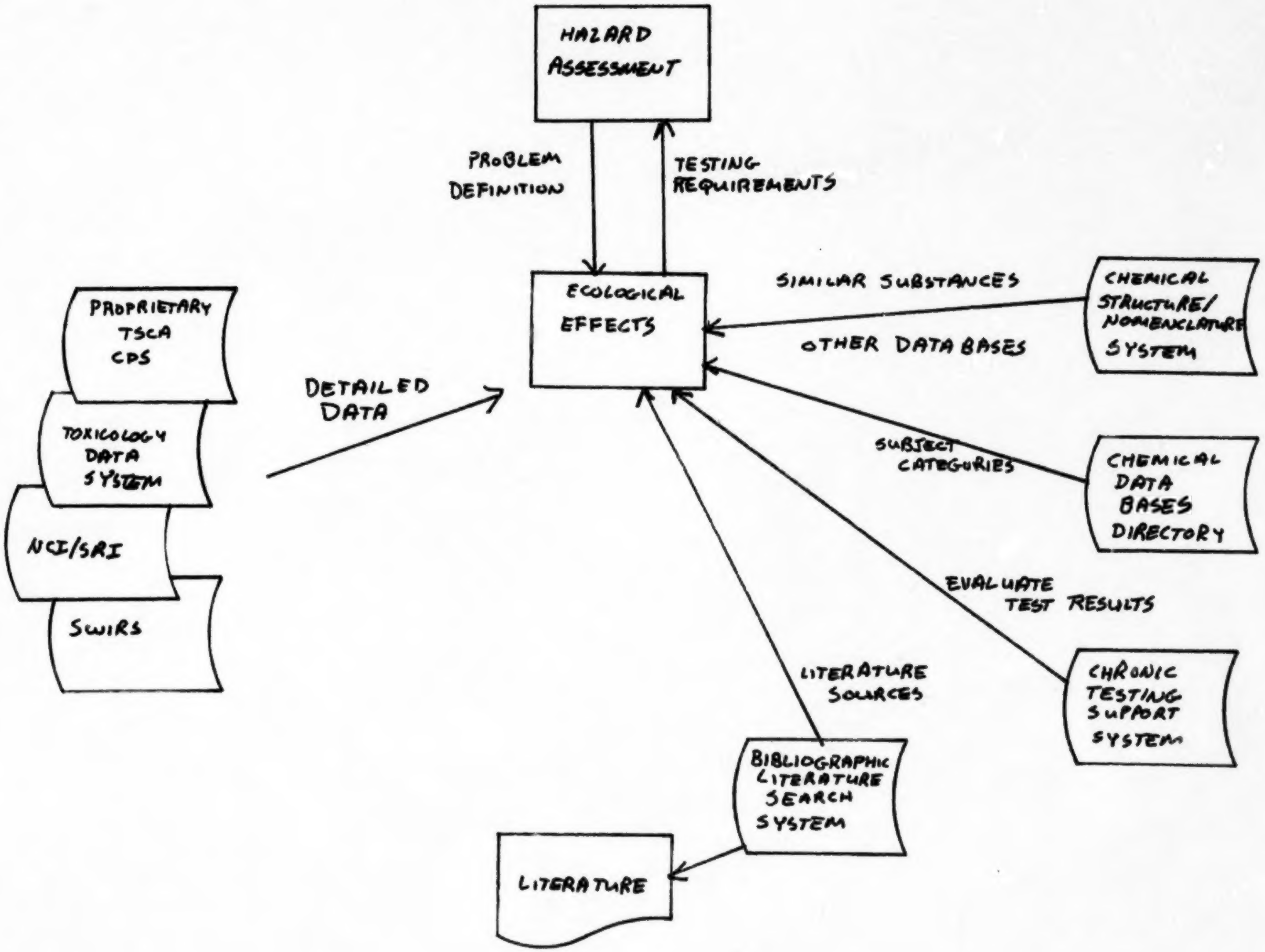
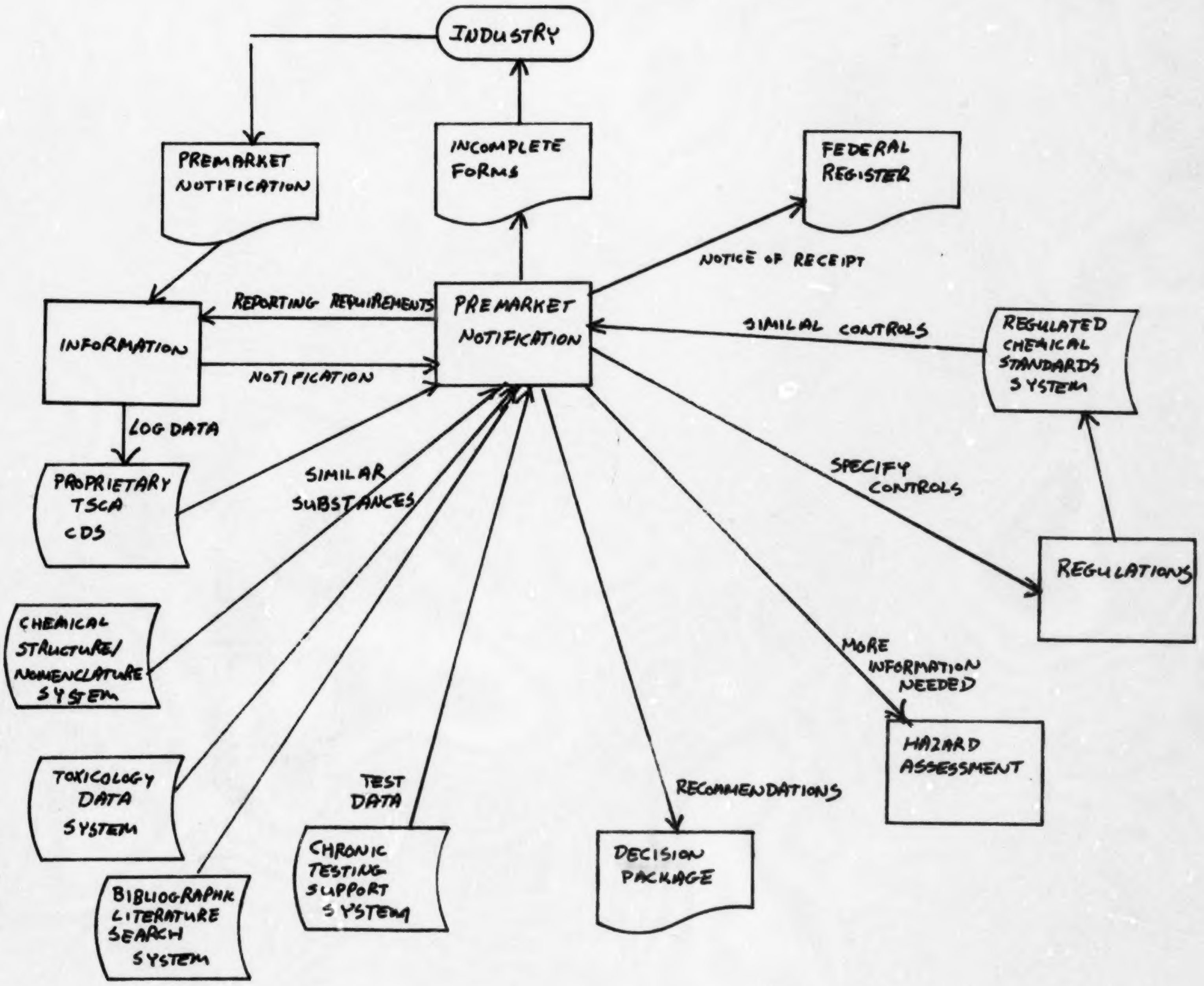

HAZARD ASSESSMENT
PROBLEM DEFINITION
TESTING REQUIREMENTS
ECOLOGICAL EFFECTS
SIMILAR SUBSTANCES
OTHER DATA BASES
CHEMICAL STRUCTURE/ NOMENCLATURE SYSTEM
SUBJECT CATEGORIES
CHEMICAL DATA BASES DIRECTORY
EVALUATE TEST RESULTS
CHRONIC TESTING SUPPORT SYSTEM
LITERATURE SOURCES
BIBLIOGRAPHIC LITERATURE SEARCH SYSTEM
LITERATURE
DETAILED DATA
PROPRIETARY TSCA CPS
TOXICOLOGY DATA SYSTEM
NCI/SRI
SWIRS

FIGURE 5-14

OTS PREMARKET NOTIFICATION GROUP


INDUSTRY
PREMARKET NOTIFICATION
INCOMPLETE FORMS
FEDERAL REGISTER
NOTICE OF RECEIPT
REPORTING REQUIREMENTS
PREMARKET NOTIFICATION
INFORMATION
NOTIFICATION
SIMILIAL CONTROLS
REGULATED CHEMICAL STANDARDS SYSTEM
LOG DATA
PROPRIETARY TSCA CDS
SIMILAR SUBSTANCES
SPECIFY CONTROLS
REGULATIONS
CHEMICAL STRUCTURE/ NOMENCLATURE SYSTEM
MORE INFORMATION NEEDED
HAZARD ASSESSMENT
TOXICOLOGY DATA SYSTEM
TEST DATA
RECOMMENDATIONS
CHRONIC TESTING SUPPORT SYSTEM
DECISION PACKAGE
BIBLIOGRAPHIC LITERATURE SEARCH SYSTEM

TSCA provides for the establishment of a notification mechanism that requires at least a 90 day notice before a new chemical can be manufactured. This time period would allow OTS to determine whether the chemical poses any risk to human health or the environment. In order to determine what chemicals are considered new, an inventory of existing chemicals must be compiled. This task will be completed in 1978 and the list of chemical substances will be stored in the Chemicals In Commerce file of the TSCA Chemical Data System. Any chemical not on the list would be considered a new chemical.

In order to determine the potential risk of a new chemical, adequate data would have to be collected and reviewed. The Premarket Notification group would develop premarket review guidelines specifying minimum data requirements or testing requirements for categories of chemicals. The Information group would generate the appropriate reporting forms and requirements needed to collect the information.

Information that may be required for premarket notification includes:

- Production, uses, byproducts and impurities
- Exposure data
- Health test data
- Environmental test data

Industry would be responsible for conducting appropriate tests, collecting necessary data, and presenting their own risk assessment for each new chemical that is to be manufactured commercially. As an alternative to conducting their own tests, industry may reference tests conducted on the chemical substance by other parties. The premarket notifications completed by industry would be submitted to the Information group who would log pertinent data in the TSCA Chemical Data System. The report submitted by industry would then be forwarded to the Premarket Notification group who would conduct a thorough review and analysis.

Since the Premarket Notification Group would be concerned with new chemicals, very little existing data would be available on the various chemical data bases. The review process would concentrate mainly on the supporting data submitted by industry.

When appropriate, data on similar chemical substances may be reviewed to determine if parallel types of risks could be involved. The Chemical Structure/Nomenclature System could be accessed to identify these similar chemicals. Once similar chemicals have been identified, the TSCA CDS could be accessed to determine exposure, adverse reactions and health study data. The Toxicology Data System could be accessed for toxicity data. Tests conducted on these similar compounds could be reviewed from data retrieved from the Chronic Testing Support System. The Regulated Chemical Standards System would yield information concerning the control of these similar compounds. If additional data is needed, the Bibliographic Literature Search System could be accessed to identify relevant journal articles pertaining to these chemicals.

Based on the premarket notification review, recommendations would be developed and documented in a decision package. The following alternatives may be included in the recommendations regarding new chemicals

- No control is necessary
- Manufacture is delayed pending further investigations
- Regulation is required
- No control is necessary but future reporting is required.

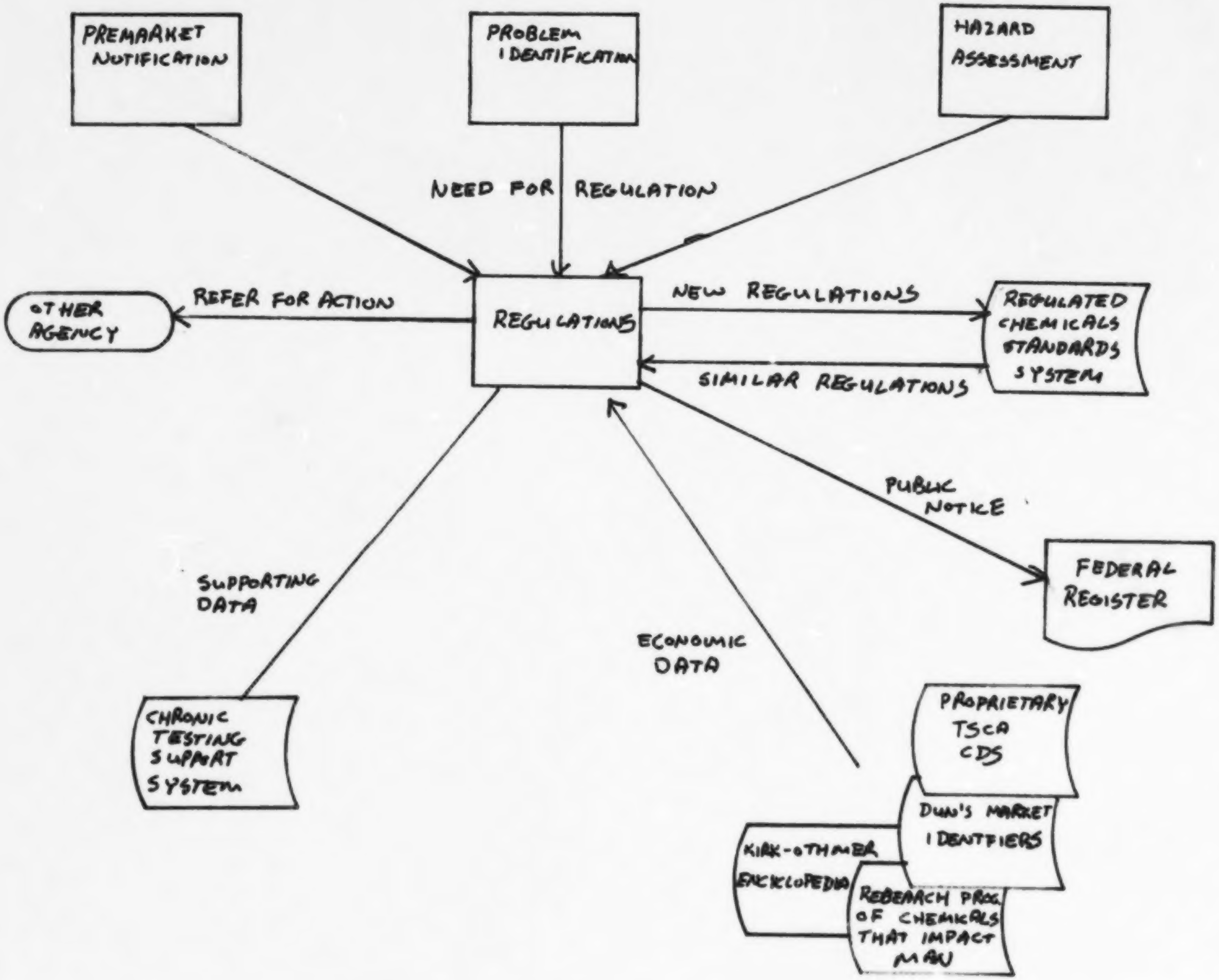
If the Premarket Notification group decides that further investigation is necessary to conduct an in-depth review, the Hazard Assessment group would specify additional tests to be conducted and developed an appropriate risk assessment. If control is required, the Regulations group would conduct a technological and economic review to determine the actual regulations to be imposed on the new chemical.

5.3.1.8 Regulations

The Regulations group has a dual prupose in controlling chemical substances. It would be responsible for preventing possible hazards and correcting existing problems through the regulatory authorities under TSCA. Figure 5-15 illustrates some of the activities associated with the Regulations group.

FIGURE 5-15

OTS REGULATIONS GROUP


PREMARKET NOTIFICATION
PROBLEM IDENTIFICATION
HAZARD ASSESSMENT
NEED FOR REGULATION
OTHER AGENCY
REFER FOR ACTION
REGULATIONS
NEW REGULATIONS
REGULATED CHEMICALS STANDARDS SYSTEM
SIMILAR REGULATIONS
PUBLIC NOTICE
FEDERAL REGISTER
SUPPORTING DATA
CHRONIC TESTING SUPPORT SYSTEM
ECONOMIC DATA
PROPRIETARY TSCA CDS
DUN'S MARKET IDENTFIERS
KIRK-OTHMER ENCKLOPEDIA
REBEARCH PROG OF CHEMICALS THAT IMPACT MAN

The need for regulation or control would be identified by three other OTS groups: the Premarket Notification group, the Hazard Assessment group, and the Problem Identification group. The Premarket Notification group would be responsible for new chemicals entering Commerce, the Problem Identification group would be responsible for existing chemicals related to notifications of risk, and the Hazard Assessment group would be responsible for recommending controls on priority chemicals selected by the Interagency Testing Committee and found to pose substantial risk. Each of these groups would submit a risk assessment to Regulations in the form of a decision package. The decision package would specify recommendations as to the types of controls that should be employed. It would be the responsibility of this group to promulgate the actual regulation.

In order to provide a repository of information dealing with environmental regulations and guidelines concerning chemical substances, the Regulations group would have a key role in the design and development of the Regulated Chemicals Standards System. The data base would contain information about the regulatory actions taken by OTS concerning specific substances or categories of chemicals and the litigation associated with these regulations. This would provide a tracking procedure that would indicate how effective the regulations is in preventing or solving and environmental hazard. The data would be accessed in an on-line interactive mode and be available for use by OTS and the general public as well.

The recommendations for controls received from other groups would mainly be the results of hazard assessments or source assessments. However, before these controls could be transformed into effective guidelines or regulations, the Regulations groups would conduct an evaluation of the economic, technological, and social impacts that control of the chemical substance might have. Data bases that might be accessed to obtain this needed information include the TSCA CDS, Dun's Market Identifiers, Mineral Commodity Survey, Research Program of Chemicals that Impact Man (SRI), and the Kirk-Othmer Encyclopedia. The Regulated Chemicals Standards System would be accessed to determine what effect regulation of similar chemicals has had on the environment and the general public.

Before issuing regulations and guidelines, the Chronic Testing Support System would be accessed to determine if enough test data exists to support the actions taken by the Regulations group. The Regulations Group may share the responsibility of taking regulatory action with other agencies or programs such as FDA.

in this case, the data obtained by OTS would be shared with other agencies in order to develop the most effective regulation or guideline. Once the regulations or guidelines have been determined by the Regulations group, they would be published in the Federal Register and appropriate information would be entered into the Regulated Chemicals Standards System. Some of the actions that could be taken by the Regulations group include:

- Banning or limiting manufacture, processing, distributing or use of a chemical
- Specifying warning labels
- Specifying disposal methods
- Specifying control measures during the manufacture process

5.3.2 Short-Term Network Description

This section discusses the data base network system linkages that would prevail in the short-term in order for OTS to fulfill its information acquisition requirements under TSCA. Short-term in this situation means approximately five years. Figure 5-16 illustrates a conceptual view of the integrated data base system network. Only the core components of the proposed Chemical Substances Information Network are included in the diagram, since these components will supply the majority of information needed by OTS. The diagram shows the stage of development the various data bases are currently in or will be in over the next five years.

As Figure 5-16 suggests, the core components of CSIN would be distributed over several data base systems. Access to the various computerized data bases would be through a dial-up mechanism except for the propriety TSCA Chemical Data System which may be directly linked to the OTS offices. Under this short-term system network, the user would not be able to switch between data bases during a particular terminal session without repeatedly going through the dial-up procedure. The one exception to this situation is the MEDLARS system at NLM. This data base system contains several CSIN core components (MEDLINE, CANCERLINE, TOXLINE, CHEMLINE, Registry of Toxic Effects, TDB, LADB, and possibly the CDBD) under the control of a single executive operating system. Once a user is connected to MEDLARS, he/she can conduct searches on any or all of the data bases without having to go through a dial-up procedure.

FIGURE 5-16

SHORT-TERM DATA BASE NETWORK

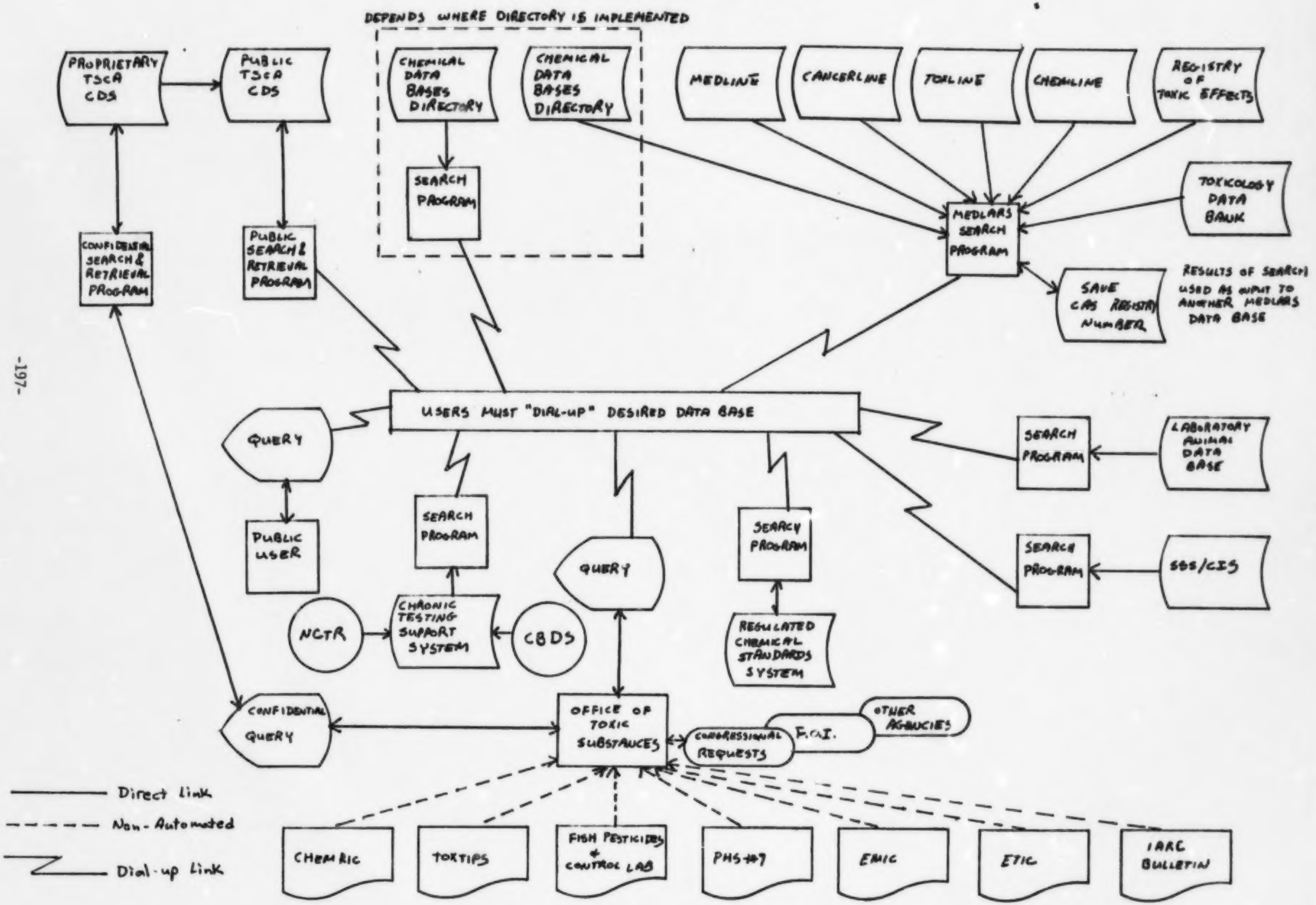
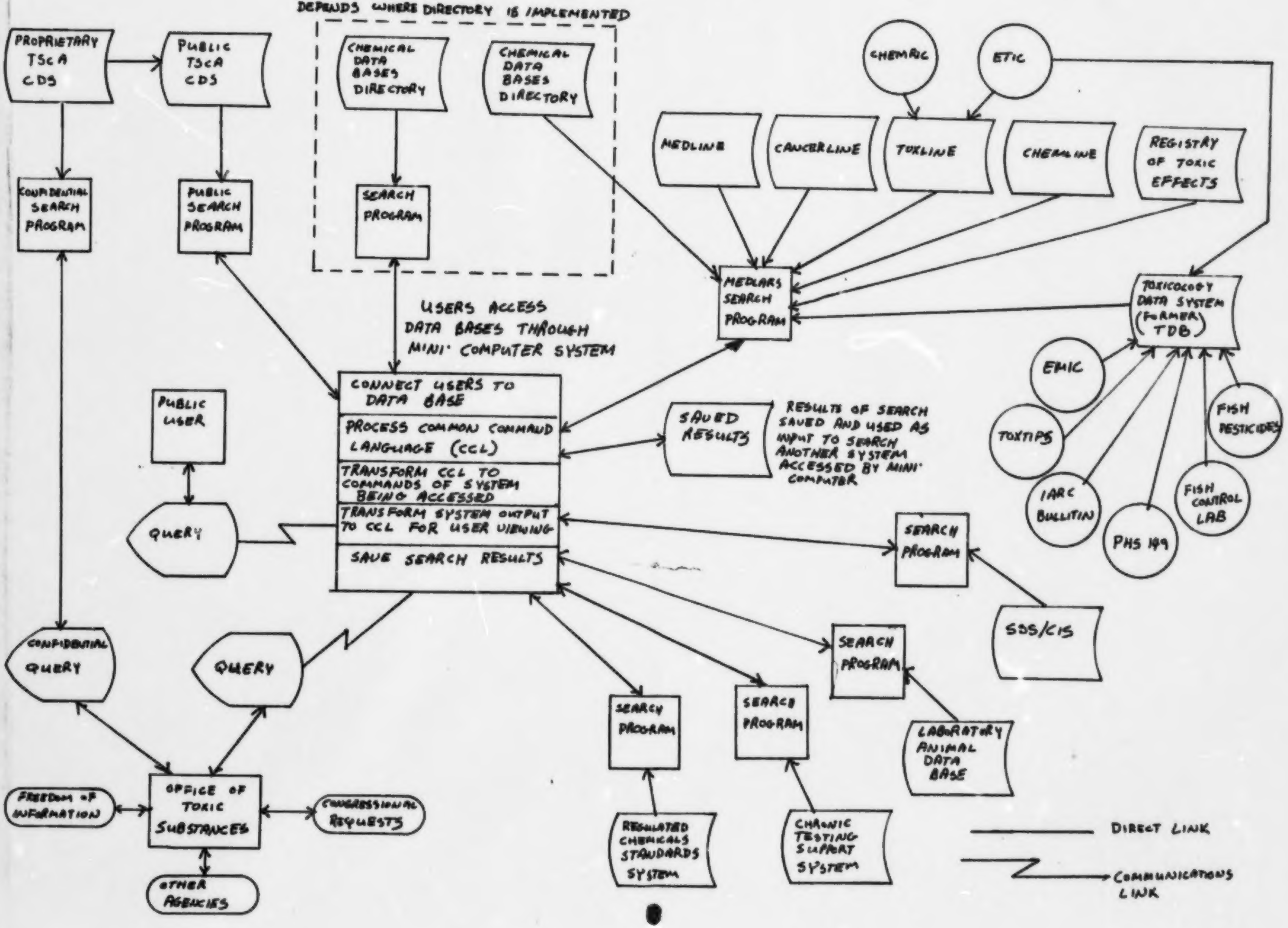

DEPENDS WHERE DIRECTORY IS IMPLEMENTED
PROPRIETARY TSCA CDS
PUBLIC TSCA CDS
CHEMICAL DATA BASES DIRECTORY
CHEMICAL DATA BASES DIRECTORY
SEARCH PROGRAM
MEDLINE
CANCERLINE
TOXLINE
CHEMLINE
REGISTRY OF TOXIC EFFECTS
TOXICOLOGY DATA BANK
MEDLARS SEARCH PROGRAM
SAVE CAS REGISTRY NUMBER
RESULTS OF SEARCH USED AS INPUT TO ANOTHER MEDLARS DATA BASE
CONFIDENTIAL SEARCH & RETRIEVAL PROGRAM
PUBLIC SEARCH & RETRIEVAL PROGRAM
USERS MUST "DIAL-UP" DESIRED DATA BASE
QUERY
PUBLIC USER
SEARCH PROGRAM
NCTR
CHRONIC TESTING SUPPORT SYSTEM
CBDS
QUERY
SEARCH PROGRAM
REGULATED CHEMICAL STANDARDS SYSTEM
SEARCH PROGRAM
LABORATORY ANIMAL DATA BASE
SEARCH PROGRAM
SBS/CIS
CONFIDENTIAL QUERY
OFFICE OF TOXIC SUBSTANCES
CONGRESSIONAL REQUESTS
F.O.I.
OTHER AGENCIES
Direct link
Non-Automated
Dial-up Link
CHEMRIC
TOXTIPS
FISH PESTICIDES & CONTROL LAB
PHS #7
EMIC
ETIC
IARC BULLETIN

FIGURE 5-17

LONG-TERM DATA BASE NETWORK


DEPENDS WHERE DIRECTORY IS IMPLEMENTED
PROPRIETARY TScA CDS
PUBLIC TScA CDS
CHEMICAL DATA BASES DIRECTORY
CHEMICAL DATA BASES DIRECTORY
CHEMRIC
ETIC
MEDLINE
CANCERLINE
TOXLINE
CHEMLINE
REGISTRY OF TOXIC EFFECTS
CONFIDENTIAL SEARCH PROGRAM
PUBLIC SEARCH PROGRAM
SEARCH PROGRAM
MEDLARS SEARCH PROGRAM
TOXICOLOGY DATA SYSTEM (FORMER) TDB
USERS ACCESS DATA BASES THROUGH MINI· COMPUTER SYSTEM
EMIC
TOXTIPS
IARC BULLITIN
PHS 149
FISH CONTROL LAB
FISH PESTICIDES
PUBLIC USER
CONNECT USERS TO DATA BASE
PROCESS COMMON COMMAND LANGUAGE (CCL)
TRANSFORM CCL TO COMMANDS OF SYSTEM BEING ACCESSED
TRANSFORM SYSTEM OUTPUT TO CCL FOR USER VIEWING
SAVE SEARCH RESULTS
SAVED RESULTS
RESULTS OF SEARCH SAVED AND USED AS INPUT TO SEARCH ANOTHER SYSTEM ACCESSED BY MINI· COMPUTER
QUERY
SEARCH PROGRAM
SDS/CIS
CONFIDENTIAL QUERY
QUERY
SEARCH PROGRAM
SEARCH PROGRAM
SEARCH PROGRAM
LABORATORY ANIMAL DATA BASE
FREEDOM OF INFORMATION
OFFICE OF TOXIC SUBSTANCES
CONGRESSIONAL REQUESTS
OTHER AGENCIES
REGULATED CHEMICALS STANDARDS SYSTEM
CHRONIC TESTING SUPPORT SYSTEM
DIRECT LINK
COMMUNICATIONS LINK

Although a user can switch access to different data bases, the results of a search on one data base cannot be used as input when searching a different data base. Since the CAS Registry Number is being considered as a possible universal identifier for chemical data bases, it is recommended that the MEDLARS system implement a procedure whereby the CAS Registry Number(s) obtained as a result of a search would be saved in a temporary file and could be used as input when searching another data base.

In order to obtain information from the core components under this network configuration, a user must be acquainted with many different search and retrieval languages and strategies. In addition, the structure and organization of the various data bases must be known in order to extract the most meaningful information. Each SEARCH PROGRAM illustrated in Figure 5-16 represents a potential procedure that a user may have to learn. The Mitre report indicated that users preferred to access a data base themselves instead of going through an information specialist or librarian. This is especially true in the case of OTS where users will conducting searches in which the results may be unanticipated and these results would be used to conduct further more detailed searches. Communicating with an information specialist in this case would be very time consuming. In order to make it easier for users to search these data bases, it is recommended that new systems developed use a common type searching language as much as possible, This common command language (CCL) would standardize as much as possible the syntax, format, construction of a search query, and response to a query for each on-line system. The CCL should be simple to understand and learn so that users inexperienced in computer programming or general computer usage could effectively intereact with the on-line data base system. This feature should be implemented when developing new CSIN data bases such as the TSCA Chemical Data System, the Chemical Data Bases Directory, the Regulated Chemicals Standards System, and the Chronic Testing Support System.

5.3.4 Long-Term Network Description

This section discusses recommendations for long-term data base network system linkages that would provide rapid and efficient retrieval of TSCA related information by OTS. In this case, long-term means that the network would take 5 to 10 years to evolve. Figure 5-17 illustrates a conceptual view of the proposed network. This network would be a further enhancement of the network described in the previous section, and would be made up of the same Chemical Substances Information Network core components that were included in the short-term network.

In the long-term version of the network, all data bases would be accessed through on-line procedures. The non-automated data bases of the short-term network would be integrated into existing computerized data bases. CHEMRIC and ETIC would be integrated into the TOXLINE data base. The current Toxicology Data Bank would become the Toxicology Data System and include automated information from ETIC, EMIC, TOXTIPS, IARC Bulletin, PHS-149, Fish Pesticides and Fish Control Lab.

The major difference between the long-term network configuration and the short-term network configuration is that users will be able to switch between systems without having to go through multiple dial-up procedures, and users will be able to access the various data bases using a single search and retrieval language. In order to provide this capability, Sigma Data recommends implementing a minicomputer based system that provides the proper communications linkage to the various data bases in the network and translates a single search and retrieval language specified by the user into the search and retrieval language of the particular data base being accessed.

This concept is feasible and is within the state-of-the-art as far as technology is concerned. The recommendations for this network concept were derived from experimental work done by the Massachusetts Institute of Technology in 1976 through a grant from the National Science Foundation. The experiment is titled The Networking of Interactive Bibliographic Retrieval Systems (NTIS # PB-252 407). Under this experiment, a working model was developed that provided the communications linkages required for users to access four systems: the MIT Intrex System, the Lockheed DIALOG System, the SDC ORBIT System, and the NLM MEDLINE System. A common command language (CCL) was developed to specify search and retrievals. The model translated the CCL into the search and retrieval commands required by the system that was being accessed. System responses were interpreted and translated into the common command language format for user viewing. The model does perform any searches, but acts as a translator of messages transmitted to and received from the data base system conducting the actual search. The actual procedures required to obtain the information are transparent to the user.

This concept solves many of the problems inherent in accessing and retrieving information from heterogeneous data bases. A user would only have to learn a single english-like search and retrieval language. This could prove important as more and more data bases are added to the network. New data bases might employ the common command language in order to eliminate the translation process. Under this minicomputer based concept, search results from one computer system could be saved in a temporary file and used as input when search a data base on another computer system. A user could access multiple computer systems without having to go through dial-up procedures every-time a switch was made.

Although the model was operational, there were some problems that would have to be overcome in a fully operational environment. Since each search and retrieval program has its own idiosyncricies, the common command language could not always accommodate these peculiarities. However, this could be overcome by enabling the user to bypass the CCL and use the actual retrieval language of the data base being accessed or developing more sophisticated software to analyze and interpret these unique situations.

It is recommended that a feasibility study be conducted within a few years to analyze this concept in more detail to determine its possible implementation into the Chemical Substances Information Network.